Extension of Fractional Imputation to General Missingness Patterns using Maximum Likelihood

Minhui Paik, Michael D. Larsen
Iowa State University, Department of Statistics, Snedecor Hall, Ames, IA 50014

Abstract

Surveys frequently have missing values for some variables for some units. Imputation is a widely used method in sample surveys as a method of handling missing data problems. We provide a new imputation procedure for various imputation models retaining many of the desirable properties of model-based imputation estimation and hot-deck imputation under fractional imputation. The main objective of this procedure is to construct an easy-to-use data set for general purpose estimation. We provide an extension of fractional imputation methods to general patterns of missing data via maximum likelihood calibration.

Key Words: EM algorithm; Missing data; Multivariate normal; Replicate variance estimation; Superpopulation model

1. The EM Algorithm for Missing Data

When the full data model is correct and the response mechanism is ignorable, the observed-data likelihood contains all relevant information about the parameters. Maximum likelihood estimates can be found by solving the estimating equations produced by setting the derivaties of the observed data log likelihood equations to zero. In some cases, the expressions for the first derivatives of the observed data log likelihood equation set to zero do not have a closed-form solution. In such a case, iterative methods can be applied. The Newton Raphson algorithm is one of the candidate algortihms for solving this problem. This method requires calculating the matrix of second partial derivativs of the observed data log likelihood function. In pratice, the method requires careful algebraic manipulations and efficient programming. An alternative strategy for incomplete-data problems, which does not require second derivatives to be calculated, is the Expectation-Maximization (EM) algorithm propsed by Dempster, Laird and Rubin (1977).

Each iteration of the EM algorithm consists of two process: the E-step (Expectaion) and the M-step (Maximization). In the E step, the functions of the missing data in the complete data log likelihood function are estimated by their conditional expectations given the observed data and the current estimated parameters. In the M-step, the completed log likelihood is maximized as it would be for ordinary ML estimation from the complete data log likelihood under the assumption that the estimate of the missing data functions from the E-step have their estimated values. A calculation involving Jensen's Inequality shows that the algorithm is guaranteed to increase the observed data likelihood at each iteration. If the log likelihood function is concave, then convergence is assured.

2. Fractional Imputation and Maximum Likelihood

A limitation so far of fractional imputation methods has been the existence of missing data in only a single variable. Here we consider multivariate normal data with arbitrary patterns of missing data in several variables. Our objective is to provie compeleted data sets with donors and weight sets retaining many of the desirable properties of ML estimation. That is, we want the weighted data set with multiple donors for each missing value to match the results for ML estimation.

In Section 4, we define the imputation method using adjusted fractional weights under the multivariate normal model and an ignorable response model, which leads to missing data. The resulting estimates of parameters using the completed data set including the imputed data will be algebraically the same as the maximum likelihood estimates using only the observed data. As with other fractional imputation methods, one can estimate other parameter that were not included in the imputation model, such as domain means and proportions. Experience suggests that the fractional imputation methods provide reasonable estimates for these other parameters.

3. Notation and Model

Consider a finite population $U = \{1, 2, \ldots, N\}$ with p variables potentially recorded for each subject. For the i^{th} element of the finite population, $y_i = (y_{i1}, \ldots, y_{ip})$, are the values of the p variables. We assume that the finite population is a random sample from a superpopulation model. In this case, we assume the superpopulation model is the p-variate normal distribution with mean μ and covariance matrix $\Sigma = (\sigma_{jk})$. For $i = 1, \ldots, n$, independently and identically distributed, $y_i \sim N_p(\mu, \Sigma)$. Let $Y_N = (y_1, y_2, \ldots, y_N)^{'}$ be the finite population of size N. The interesting parameters are the finite population mean of each variable and the finite population covariance between two variables. There are defined as follows: $\bar{Y}_{j,U} = \frac{1}{N}\sum_{i\in U} y_{ij}$ and $S_{jk,U} = \frac{1}{N-1}\sum_{i\in U}(y_{ij} - \bar{Y}_{j,U})(y_{ik} - \bar{Y}_{k,U})$. For simplicity, the shorter notation $Y_j$and S_{jk} are used in this section. Note that S_{jk} can be expressed as functions of the population means $\bar{Y}_j = N^{-1}\sum_{i\in U} y_{ij}$, $\bar{Y}_k = N^{-1}\sum_{i\in U} y_{ik}$, $\bar{Y}_{jj} = N^{-1}\sum_{i\in U} y_{ij}^2$, $\bar{Y}_{kk} = N^{-1}\sum_{i\in U} y_{ik}^2$, and $\bar{Y}_{jk} = N^{-1}\sum_{i\in U} y_{ij}y_{ik}$. as $S_{jk} = \frac{N}{N-1}(\bar{Y}_{jk} - \bar{Y}_j\bar{Y}_k)$.

If data were observed on the complete sample, estimators of $\bar{Y}_j$ and $\bar{Y}_{jk}$ based on the sample A with size n are $\bar{y}_{j,n} = \sum_{i\in A} w_i y_{ij}$ and $\bar{y}_{jk,n} = \sum_{i\in A} w_i y_{ij} y_{ik}$, where $w_i = N^{-1}\pi_i^{-1}$, $\pi_i = P(i \in A)$ is the probability that unit i is in the sample, and A is the set of indices in the sample.

An estimator of S_{jk} is constructed as follows: $s_{jk,n} = \frac{N}{N-1}(\bar{y}_{jk,n} - \bar{y}_{j,n}\bar{y}_{k,n})$, where $\bar{y}_{jk,n}$, $\bar{y}_{j,n}$, and $\bar{y}_{k,n}$ are sample estimates of the corresponding population means. For simplicity, the shorter notation y_j, y_{jk} and s_{jk} are used in the remainder of this section. By the definition of w_i, we have $E(\bar{y}_j|F_N) = \bar{Y}_j$, $E(\bar{y}_{jj}|F_N) = \bar{Y}_{jj}$, $E(\bar{y}_{jk}|F_N) = \bar{Y}_{jk}$, and $E(s_{jk}|F_N) = S_{jk}$, where $F_N = \{y_1, \ldots, y_N\}$.

Assuming that π_i is greater than zero for all i and does not depend on values of y for any units in the population, since the population comes from a model for which all moments exist, the folllowing lemma is true.

LEMMA: Under the superpopulation model and an ignorable sampling mechanism, $\bar{y}_j$ and s_{jk} are consistent estimators for finite population parameters.

4. Missing Data and the Proposed Method

Let Y_n be a sample from the finite population produced by the sampling design. Assume that the sample design and the response mechanism are ignorable. That is, we assume that π_i is greater than zero for all i and does not depend on values of y for any units in the population. We also assume the probability that a sampled unit is observed does not depend on unobserved variables. We write $Y_n = (Y_{obs}, Y_{mis})$, where Y_{obs} is the set of observed values and Y_{mis} is the set of missing values in the sample. Under a missing at random (MAR) assumption, treating the sample as an *iid* sample from a multivariate normal distribution, the marginal distribution of the observed data Y_{obs} can be used to construct the correct likelihood for use in estimating the model parameters. In the multivariate normal case, under the MAR assumption, the ML estimates of μ and Σ can be obtained by maximizing the observed log likelihood with respect to μ and Σ. In some cases, however, even for the multivariate normal model, the observed log likelihood equations do not have a closed form solution. As was mentioned, iterative methods, such as Newton-Raphson, Fisher scoring, and the EM algorithm can be used to produce ML estimates.

If the only interest were to produce estimates of model parameters without respect to the finite population and its sampling design, then estimates of μ and Σ using maximum likelihood estimation would have been sufficient. The goal here, however, is estimation of finite population parameters. Further, the estimated parameters in the imputation model do not necessarily lead to estimates of other parameters not included in the model, such as domain means and proportions. To repeat, our objective in this section is to provide a method for making an easy-to-use data set for the analyst that retains properties of the ML estimates and at the same time provides reasonalbe estimates for other parameters.

To acheive our objective, the imputed values y_{ij}^* for subject i on variable j has to satisfy the following conditions. We define the response indicator variable of y_{ij} by $R_{ij} = 1$ if variable j is observed for unit i and $R_{ij} = 0$ if variable j is not observed for unit i.

When $R_{ij} = 0$, $R_{ik} = 0$, and $R_{is} = 1$ for $s \neq j, k$, $y_{ij}^* = E(y_{ij}|y_{obs,i}, \hat{\theta})$, $y_{ij}^{*2} = E(y_{ij}^2|y_{obs,i}, \hat{\theta})$, and $y_{ij}^* y_{ik}^* = E(y_{ij}y_{ik}|y_{obs,i}, \hat{\theta})$, where $y_{obs,i}$ denotes the set of variables observed for unit i and $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ are ML estimates, which possibly are obtained by iterative methods.

Since a single imputated donor value can not satisfy the above conditions, our approach is to use multiple donors and assign adjusted fracional weights to the donors in order to satisfy the conditions. The imputed values based on several donors and fractional weights can be defined as follows: $y_{ij}^* = \sum_{t\in A_{D,i}} w_{it}^* y_{tj}$, where $A_{D,i}$ is the donor set of indices for missing unit i. Note that this donor set for unit i can be constructed by a systematic sampling method from available donors sorted in some manner, simple random sampling without replacement from the observed cases, or selection of donors using some nearest neighbor method. If it is not important to use observed values as imputed values, the imputed values can be generated from the conditional distribution given by observed cases y_{obs} and ML estimates. Then, the proposed method consists of finding the fractional weights satisfying the following constraints. Two cases can be considered. First, for $R_{ij} = 0$ and $R_{ik} = 1$, $j \neq k$,

$$\sum_{t \in A_{D,i}} w^*_{it} \left(1, y_{tj}, y^2_{tj}\right) = \left(1, E(y_{ij}|y_{obs,i}, \hat{\theta}), E(y^2_{ij}|y_{obs,i}, \hat{\theta})\right). \tag{1}$$

Second, for $R_{ij} = 0$, $R_{ik} = 0$ and $R_{is} = 1$ for $s \neq j, k$, the constaints are

$$\begin{aligned} \sum_{t \in A_{D,i}} w^*_{it} \left(1, y_{tj}, y_{tk}, y_{tj}y_{tk}, y^2_{tj}, y^2_{tk}\right) &= \\ \left(1, E(y_{ij}|y_{obs,i}, \hat{\theta}), E(y_{ik}|y_{obs,i}, \hat{\theta}), E(y_{ij}y_{ik}|y_{obs,i}, \hat{\theta}), E(y^2_{ij}|y_{obs,i}, \hat{\theta}), E(y^2_{ik}|y_{obs,i}, \hat{\theta})\right)&. \end{aligned} \tag{2}$$

We can use a regression weighting technique or an empirical likelihood technique to find a solution to (1) and (2). To avoid the chance of extreme weights, including possibly negative weights, the nonnegative fractional weights method of Paik and Larsen (2007) or a modified Newton-Raphson method as in Chen, Sitter and Wu (2002) can be used to solve the constraints. The size of donor sets for missing values do not necessarily need to be very large in order to do this in general.

5. An Example: Trivariate Normal Sample with Bivariate Missing Data

Suppose that (y, z, x) have a trivariate normal distribution with a mean vector $\mu = (\mu_y, \mu_z, \mu_x)$ and a covariance matrix Σ with entries $\tilde{\sigma} = (\sigma_{yy}, \sigma_{yz}, \sigma_{yx}, \sigma_{zz}, \sigma_{zx}, \sigma_{xx})$. Let $\theta = (\mu, \tilde{\sigma})$. Suppose a random sample with a certain pattern of missing data is obtained from this distribution. The values of x are observed for all units. Some values of y and z are missing under the MAR assumption. We can define four groups of units based on their missing data patterns. The first group A_{rr} of units have both y and z observed. The second group A_{mr} has z observed but is missing y. The third group A_{rm} has y observed but is missing z. The fourth group A_{mm} has both y and z missing. Under a MAR assumption, the ML estimates $\hat{\theta}$ of θ can be obtained by maximizing the observed data log likelihood, possibly with iterative methods of solution.

The constraints (1) and (2) in this situation can be expressed as follows. For $i \in A_{mr}$, $\sum_{t \in A_{D,i}} w^*_{it}(1, y_t, y^2_t) = \left(1, E(y_i|z_i, x_i, \hat{\theta}), E(y^2_i|z_i, x_i, \hat{\theta})\right)$. For $i \in A_{rm}$, $\sum_{t \in A_{D,i}} w^*_{it}(1, z_t, z^2_t) = \left(1, E(z_i|y_i, x_i, \hat{\theta}), E(z^2_i|y_i, x_i, \hat{\theta})\right)$. For $i \in A_{mm}$, $\sum_{t \in A_{D,i}} w^*_{it}(1, y_t, z_t, y_t z_t, y^2_t, z^2_t) = \left(1, E(y_i|x_i, \hat{\theta}), E(z_i|x_i, \hat{\theta}), E(y_i z_i|x_i, \hat{\theta}), E(y^2_i|x_i, \hat{\theta}), E(z^2_i|x_i, \hat{\theta})\right)$.

Since the data are assumed to come from a multivariate normal distribution, we can give explicit formulas for expectations and conditional expectations:

$$\begin{aligned} E(y_i|z_i, x_i, \hat{\theta}) &= \hat{\mu}_y + (\hat{\sigma}_{yz}, \hat{\sigma}_{yx}) \begin{pmatrix} \hat{\sigma}_{zz} & \hat{\sigma}_{zx} \\ \hat{\sigma}_{zx} & \hat{\sigma}_{xx} \end{pmatrix}^{-1} \begin{pmatrix} z_i - \hat{\mu}_z \\ x_i - \hat{\mu}_x \end{pmatrix}, \\ V(y_i|z_i, x_i, \hat{\theta}) &= \hat{\sigma}_{yy} - (\hat{\sigma}_{yz}, \hat{\sigma}_{yx}) \begin{pmatrix} \hat{\sigma}_{zz} & \hat{\sigma}_{zx} \\ \hat{\sigma}_{zx} & \hat{\sigma}_{xx} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\sigma}_{yz} \\ \hat{\sigma}_{yx} \end{pmatrix}, \\ E(y_i|x_i, \hat{\theta}) &= \hat{\mu}_y + \frac{\hat{\sigma}_{yx}}{\hat{\sigma}_{xx}}(x_i - \hat{\mu}_x), \\ V(y_i|x_i, \hat{\theta}) &= \hat{\sigma}_{yy} - \frac{\hat{\sigma}^2_{yx}}{\hat{\sigma}_{xx}}, \quad \text{and} \\ C(y_i, z_i|x_i, \hat{\theta}) &= \hat{\sigma}_{zy} - \frac{\hat{\sigma}_{yx}\hat{\sigma}_{zx}}{\hat{\sigma}_{xx}}. \end{aligned} \tag{3}$$

Similarly, we can calculate other conditional expectations $E(z_i|x_i, \hat{\theta})$, $V(z_i|x_i, \hat{\theta})$, $E(z_i|y_i, x_i, \hat{\theta})$, and $V(z_i|y_i, x_i, \hat{\theta})$.

The proposed imputed estimators of population means in Section 3 are defined respectively as

$$\bar{y}_{I,j} = \sum_{i \in A} \sum_{t \in A_{D,i}} w_i w^*_{it} y_{tj} \quad \text{and} \quad \bar{y}_{I,jk} = \sum_{i \in A} \sum_{t \in A_{D,i}} w_i w^*_{it} y_{tj} y_{tk}, \tag{4}$$

where $w_{it} = 0$ when $R_{ij} = 1$ and $i \neq t$, $A_{D,i}$ has only one unit y_{ij}, and $w_{ii} = 1$. In addition, the imputed estimators of population covariance in Section 3can be written as

$$s_{I,jk} \quad = \quad \frac{N}{N-1}(\bar{y}_{I,jk} - \bar{y}_{I,j}\bar{y}_{I,n}). \tag{5}$$

6. A Theoretical Result

THEOREM: The imputed estimators in (4) and (5) based on the fractional imputation described in this section are consistent estimators for finite population parameters under the superpopulation model in Section 3 and the following assumptions. First, assume that

$$K_1 \leq \frac{n}{N} w_i \leq K_2 \tag{6}$$

for all $i = 1, \ldots, N$, uniformly in n, where K_1 and K_2 are fixed constants implying that no extreme weights dominate the others. Second, suppose that the maximum likelihood estimator $\hat{\theta}$ of θ is available and, under some regularity conditions, $\hat{\theta} = \theta + O_p(1/\sqrt{n})$. Third, the first derivatives of the conditinal distribution $E(y_{mis}|y_{obs}, \theta)$ are bounded.

The proof of this result will be included a paper to be submitted to a refereed journal. Please contact the authors if you are interested.

7. Discussion of Practical Issues

In order to use the proposed fractional imputation method, one must construct weights satisfying (1) and (2). It is important to avoid the extreme weights beacuse applying these weights to make estimates for various domain means and proportions may produce unrealistic estimates for some domains and proportions. In this section, we consider the method of constructing fractional weights to avoid the chance of extreme weights like negative weights.

We want to select fractional weights w^*_{tj} satisfying (1) and (2) with $0 \leq w^*_{tj} \leq 1$ for $t \in A_{D,i}$. This leads to a constrained minimization problem that can be solved by Lagrange multipliers. We must minimize the following expression. For each missing unit i,

$$Q(w^*_{tj}) \quad = \quad d(w^*_{tj}, \alpha_{tj}) - \lambda^{'} T(w^*) - \lambda_0 \left(\sum_{t \in A_{D,i}} w^*_{tj} - 1 \right), \tag{7}$$

where $\alpha_{tj} = 1/M$ is an initial weight under simple random sampling without replacement, M is the size of donor set, $T(w^*)$ is a re-expression of the statistic in terms of the fractional weights, $w^* = \{w^*_{tj}, t \in A_{D,i}\}$ and $d(.,.)$ is a distance measure. Note that initial weights α can be considered the empirical probabilites on the donor variables.

Various distance measures can be used for our problem. Specially, Hellinger distance and Entropy distance measures can be applied for nonnegative weights. A common distance measure between two sets of probabilities is Entropy measure, $d(w^*_{tj}, \alpha_{tj}) = \sum_{t \in A_{D,i}} w^*_{tj} log(\frac{w^*_{tj}}{\alpha_{tj}})$. Then we need to solve the following expression under entropy distance measure:

$$\log(Mw^*_{ij}) + 1 - \lambda^{'} \frac{\partial}{\partial w^*_{ij}} T(w^*) - \lambda_0 = 0, \tag{8}$$

subject to the constraints $T(w^*) = 0$ and $\sum_{tinA_{D,j}} w^*_{tj} = 1$.

In the case of (1), implying that only j^{th} variable is missing among p variables, $T(w^*)$ can be written as

$$T(w^*) = \sum_{t \in A_{D,i}} w^*_{it} \left(1, y_{tj} - E(y_{ij}|y_{obs,i}, \hat{\theta}), y^2_{tj} - E(y^2_{ij}|y_{obs,i}, \hat{\theta})\right) \tag{9}$$

and the expression in (8) can be reduced to $\log(Mw^*_{tj}) + 1 - \lambda_1 y_{tj} - \lambda_2 y^2_{tj} - \lambda_0 = 0$. Using the previous formula and $\sum_{t \in A_{D,i}} w^*_{tj} = 1$, the adjusted fractional weights can be written as $w^*_{tj} = \frac{e^{\hat{\lambda}_1 y_{tj} + \hat{\lambda}_2 y^2_{tj}}}{\sum_{t \in A_{D,i}} e^{\hat{\lambda}_1 y_{tj} + \hat{\lambda}_2 y^2_{tj}}}$, where the $\hat{\lambda}_k$, $k = 1, 2$ are the solutions to $T(w^*) = 0$ in (9). In general, a Newton-Raphson method can be used to solve the nonlinear equations $T(w^*) = 0$. The resulting fractional weights will be positive and $0 \leq w^*_{tj} \leq 1$, $t \in A_{D,i}$ and be satisfying the constraints (2) or (1).

Note that the Euclidean distance always has a solution in (7) but the resulting weights can be negative and extremly large. Otherwise, the Entropy distance measure is guaranteed to obtain non-negative weights but a solution is not guaranteed for some unlucky samples of donor. To avoid an "unlucky" donor set, one apprach is to use a modified sampling mechansim to select donors where the first and second moments in donor sets are possibly close to those of the conditional distribution given by observed data and estimated values of parameters.

Further practical consideration for fractional imputation are discussed in regard to the implementation of the simulation.

8. Variance Estimation

When the final user is different than the data provider, it is common practice to include a set of replicate weights in the data set for purposes of variance estimation. Fuller and Kim (2005) point out the advantage of providing a single set of replicate weights: "A single set of replicates can be used for variance estimation for imputed variables, variables observed on all respondents, and under assumptions, for function of the two types of variables."

To consider replication variance estimation, let a replication variance estimator for the complete sample be

$$\hat{V}(\hat{\xi}_n) = \sum_{k=1}^{K} c_k(\hat{\xi}_n^{(k)} - \hat{\xi}_n)^2, \tag{10}$$

with ξ_i being any component of the matrix $y_i y_i^{'}$, $\hat{\xi}^{(k)}$ is the k-th estimate of ξ_N, based on the observation included in the k-th replicate, K is the number of replicates and c_k is a factor associated with replicate k determined by the replication method. When the original estimator $\hat{\xi}_n$ is a linear estimator, the k-th replicate estimate of $\hat{\xi}_n$ can be written as $\hat{\theta}_n^{(k)} = \sum_{i \in A} w_i^{(k)} \xi_i$, where $w_i^{(k)}$ denotes the replicate weight for the ith unit of the k replication.

Let the k^{th} replicate of the fractional imputation estimator be $\hat{\xi}_{I,n}^{(k)}$. Let a replication variance estimator for the fractional imputed estimator be

$$\hat{V}(\hat{\xi}_{I,n}) = \sum_{k=1}^{K} c_k(\hat{\xi}_{I,n}^{(k)} - \hat{\xi}_{I,n})^2, \tag{11}$$

where $\hat{\xi}_{I,n} = \sum_{i \in A} \sum_{t \in A_{D,i}} w_i w_{it}^* \xi_i$ and $\hat{\xi}_{I,n}^{(k)} = \sum_{i \in A} \sum_{t \in A_{D,i}} w_i^{(k)} w_{it}^{(*k)} \xi_i$.
The replicated fractional weights $w_{tj}^{*(k)}$ in (11) are to be constructed using a regression weighting technique that leads to a solution satisfying the following constraints. For $R_{ij} = 0$ and $R_{ik} = 1, j \neq k$, $\sum_{t \in A_{D,i}} w_{it}^* \left(1, y_{tj}, y_{tj}^2\right) = \left(1, E(y_{ij}|y_{obs,i}, \hat{\theta}^{(k)}), E(y_{ij}^2|y_{obs,i}, \hat{\theta}^{(k)})\right)$. For $R_{ij} = 0$, $R_{ik} = 0$ and $R_{is} = 1$ for $s \neq j, k$,
$\sum_{t \in A_{D,i}} w_{it}^{(k)*} \left(1, y_{tj}, y_{tk}, y_{tj}y_{tk}, y_{tj}^2, y_{tk}^2\right) = \Big(1, E(y_{ij}|y_{obs,i}, \hat{\theta}^{(k)}), E(y_{ik}|y_{obs,i}, \hat{\theta}^{(k)}), E(y_{ij}y_{ik}|y_{obs,i}, \hat{\theta}^{(k)}),$
$E(y_{ij}^2|y_{obs,i}, \hat{\theta}^{(k)}), E(y_{ik}^2|y_{obs,i}, \hat{\theta}^{(k)})\Big)$, where $\hat{\theta}^{(k)}$ is the MLE estimate of θ based on the k^{th} replicate sample.

Provided that the variance estimator of the complete estimator in (10) is consistent, the proposed variance estimator of the FI estimator is also consistent for the finite population means and covariances.

9. Simulation

In order to demonstrate the performance of the proposed estimators, we generate a finite population of size $N = 5,000$ with three variables $U_i = (Y_i, Z_i, X_i)$ from trivariate normal distribution with the mean vector $\mu = (1, 2, 3)$ and covariance matrix Σ with entries $\tilde{\sigma} = (\sigma_{yy} = 1, \sigma_{yz} = 0.8, \sigma_{yx} = 1, \sigma_{zz} = 2, \sigma_{zx} = 1.5, \sigma_{xx} = 2)$. In addition, an indicator of membership in a domain, D_i, is generated from the uniform $(0, 1)$ distribution, independent of Y_i, Z_i and X_i. The domain will be defined by D_i being below a set cutoff value.

Monte Carlo samples of size $n = 200$ were generated by simple random sampling from the finite population. From each sample, we also generated response indicator variables R_{1i} and R_{2i} from a Bernoulli distribution with the response rates $p_1 = 0.65$ and $p_2 = 0.55$, independently. The variable Y_i is observed if and only if $R_{1i} = 1$. The variable Z_i is observed if and only if $R_{2i} = 1$. The probability of responding to both variables is then $0.55 * 0.65 = 0.3575$, or 35.75%. In simulations, the average rate of responding to both variables was approximately 36.6%.

For the comparison, we used following methods:

1. ML Maximum Likelihood Estimation using EM algorithm.
2. FI Fractional Imputation Estimation proposed in this section with $M = 10$ donors.
3. MI Multiple imputation with $M = 10$ repeated imputations.

In ML, we used estimates based on complete data set (both Y and Z are observed) as the starting values.

In the case of FI, the selection of donors must be done carefully. Since the fractional weights constructed by the regression weighting technique in FI can be quite variable, producing some large weights, or even negative weights to satisfy the constraints (1) and (2). In this simulation, we used a slightly modified selection method based on the nearest neighbor criterion and simple random sampling. The nearest neighbor criterion is used for avoiding some extreme weights. Simple random sampling without replacement is used for preserving the observational distribution, instead of relying on the model to generating simulated values for imputation. In particular, for missing unit j, two closet donors are selected where one is the closest one $E(U_{mis}|U_{obs}, \hat{\theta})$ among the set of observed unit having U_{obs}-values greater than $E(U_{mis}|U_{obs}, \hat{\theta})$ and the other one is the closest one $E(U_{mis}|U_{obs}, \hat{\theta})$-value among the set of observed unit having U_{obs}-values less than $E(U_{mis}|U_{obs}, \hat{\theta})$. After selecting two donors, the $M - 2$ donors are selected with simple random sampling without replacement. One option when the fractional weights are negative is to select a new set of donors in the hope that the resulting weights will all be positive. When some of the final fractional weights w^*_{tj} are still negative or extreme, then the algorithm for producing nonnegative fractional regression weights proposed by Paik and Larsen (2007) was applied to produce nonnegative fractional weights satisfying (1) and (2).

For MI, the missing values are generated from the posterior predictive distribution of the data given the observed values. The method of multiple imputation for the multivariate normal model is used as follows:

MI 1 For each repetition of the imputation, $k = 1, \ldots, M$, draw $\Sigma^*_{(k)}|U_{obs} \sim^{i.i.d}$ Inverse-Wishart$_{v-1}(S)$, where v is the size of the set A_{rr} and S is the sum of squares matrix about the sample mean on complete data A_{rr}: $S = \sum_{A_{rr}}(U_i - \bar{U}_r)(U_i - \bar{U}_r)'$, where $\bar{U}_r$ is the mean of U_i on A_{rr}.

MI 2 Generate $\bar{U}^*_{(k)}| \left(U_{obs}, \Sigma^*_{(k)}\right) \sim^{i.i.d} N(\bar{U}_r, \Sigma^*_{(k)})$.

MI 3 For missing unit j, generate $e^*_{j(k)}| \left(\bar{U}^*_{(k)}, \Sigma^*_{(k)}\right) \sim^{i.i.d} N(0, \Sigma^*_{(k)})$. Then $U_{j(k)} = E(U_j|U_{obs}, \bar{U}^*_{(k)}) + e^*_{j(k)}$ is the values associated with unit j for k^{th} imputation.

MI 4 Repeat steps 1-3 independently M times.

10. Simulation Results

The population parameters that are studied in this simulation are listed below.

1. Population values: $\bar{Y}_N, \bar{Z}_N, S_{yy,N}, S_{zz,N} S_{yz,N}, S_{yx,_N}$ and $S_{zx,_N}$,
2. Domain values: $\bar{Y}_{D,N}$ and $\bar{Z}_{D,N}$ are means of Y and Z where $D < 0.45$,
3. $P_{y,N}$ = proportion of $Y > 1.65$, and
4. $P_{z,N}$ = proportion of $Z < 1.38$.

For variance estimation, we have considered the FI estimator and the MI estimator of variance. For the FI variance estimator, we used the jackknife variance estimation method discussed in previous section. In case of the MI variance estimator, the simple variance formula of Rubin (1987) is used.

The Monte Carlo results for 5,000 samples generated are given Table 1 and Table 2. Table 1 shows the mean and variance of the point estimators for three methods. The properties of the variance estimators (MI and FI) are given in Table 2. Table 2 shows the relative bias and t-statistic for the variance estimators. The relative bias of the variance estimator is estimated by $RB = \frac{E_{MC}(\hat{V}(\xi_I)) - V_{MC}(\hat{\xi}_I)}{V_{MC}(\hat{\xi}_I)} \times 100$, where V_{MC} is the Monte Carlo variance given in Table 1. The t-statistic is the statistic used to test the significance of the bias of the variance estimator:

$t = \sqrt{B} \times \frac{E_{MC}(\hat{V}(\xi)) - V_{MC}(\hat{\xi}_I)}{V_{MC}(\hat{V}(\xi))}$, where B is the number of replications.

The proposed FI estimator provide the same results as the EM method for the finite populations parameters except for domain means and proportions. The estimation of domain mean and proportions are not available based on EM methods.

In Table 1, the proposed FI estimator shows more efficency than the MI estimator for all parameters except the proportions. Results are about the same when the imputed values are generated from the conditional distribution given the observed data.

In Table 2, the replication variance estimation procedures are nearly unbiased for all parameters except for the domain means in this set up. Since the adjusted replicate weights constucted as part of the process for estimating the variance of the fractional imputed estimator for the finite population means was applied to obtain estimates for variance of the domain estimators, variance estimation for the domain mean estimators is slighly biased. However, the FI variance estimators for domain means are much better than the MI variance estimators. The MI variance estimation procedure provides consistent estimates for the variance of the parameter estimates in the imputation model. Even though the correct imputation model is used, the variance estimators are seriously biased for domain means and proportions which are not included in the imputation model. A bias of the MI variance estimator for domain means where the domain information is not used for imputation was pointed out by Fay (1992) and Kim and Fuller (2004).

11. Conclusion

Based on the simulation results, the proposed fractional imputation method seems to be a good imputation method because it retains the diserable propoerties of maximum likelihood estimation when estimating the parameters of the super population model, uses actually observed values, and produces a single set of general purpose replicate fractional weights. In addition, it provides reasonable estimates for other parameters that were not included in the imputation models. As with other fractional imputation methods, an easy-to-use data set was constructed for general purpose estimation. For the completed data set constructed by the proposed procedure, the standard estimates at the aggregate level of analysis are equivalent to model-based imputation estimates based on maximum likelihood for parameters in the imputation model.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 0532413. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Chen, J., Sitter, R.R., Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimator for surveys. *Biometrika*, 89, 230-237.

Dempster, A.P., Laird, N.M., and Rubin, D. B. (1977), Maximum Likelihood from Incomplete Data Via the EM Algorithm (with comments). *Journal of the Royal Statistical Society*, B39, 1-37.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research methodology section*, American Statistical Association, 227-232.

Fuller, W.A., and Kim, J.K. (2005). Hot deck imputation for the response model. *Survey Methodology*, 31, 139-149.

Kim, J.K., and Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.

Paik, M., and Larsen, M.D. (2007). Weight adjustments for fractional regression hot deck imputation. *Proceedings of the Survey Research Methods Section, ASA.*

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley-Interscience.

Table 1: Monte Carlo means and variances for imputation estimators, based on 5,000 samples.

Parameter	Method	Mean	Variance
$\bar{Y}_N$	Actual	1.02	
	EM	1.02	0.0062
	FI(M=10)	1.02	0.0062
	MI(M=10)	1.02	0.0065
$\bar{Z}_N$	Actual	2.03	
	EM	2.03	0.0137
	FI(M=10)	2.03	0.0137
	MI(M=10)	2.03	0.0147
$\bar{S}_{yy,N}$	Actual	1.00	
	EM	1.00	0.0146
	FI(M=10)	1.00	0.0146
	MI(M=10)	0.97	0.0153
$\bar{S}_{zz,N}$	Actual	2.03	
	EM	2.02	0.0063
	FI(M=10)	2.02	0.0063
	MI(M=10)	2.04	0.0087
$\bar{S}_{yz,N}$	Actual	0.82	
	EM	0.82	0.0207
	FI(M=10)	0.82	0.0207
	MI(M=10)	0.82	0.0213
$\bar{S}_{yx,N}$	Actual	1.01	
	EM	1.01	0.0172
	FI(M=10)	1.01	0.0172
	MI(M=10)	1.01	0.0177
$\bar{S}_{zx,N}$	Actual	1.51	
	EM	1.51	0.0384
	FI(M=10)	1.51	0.0384
	MI(M=10)	1.51	0.0404
$\bar{Y}_{D,N}$	Actual	1.02	
	FI(M=10)	1.02	0.0135
	MI(M=10)	1.02	0.0191
$\bar{Z}_{D,N}$	Actual	2.03	
	FI(M=10)	2.03	0.0211
	MI(M=10)	2.03	0.0245
$P_{y,N}$	Actual	0.26	
	FI(M=10)	0.26	0.0012
	MI(M=10)	0.26	0.0012
$P_{z,N}$	Actual	0.32	
	FI(M=10)	0.32	0.0013
	MI(M=10)	0.32	0.0013

Table 2: Relative biases and t-statistics for the variance estimators, based on 5,000 samples.

Parameter	Method	RB(%)	t-statistic
$\bar{Y}_N$	FI(M=10)	0.51	0.12
	MI(M=10)	-1.61	-0.80
$\bar{Z}_N$	FI(M=10)	4.06	2.05
	MI(M=10)	5.09	2.47
$\bar{S}_{yy,N}$	FI(M=10)	-2.66	-1.23
	MI(M=10)	-2.76	-1.38
$\bar{S}_{zz,N}$	FI(M=10)	2.34	1.42
	MI(M=10)	-3.07	1.48
$\bar{S}_{yz,N}$	FI(M=10)	1.07	0.37
	MI(M=10)	5.07	2.50
$\bar{S}_{yx,N}$	FI(M=10)	-1.28	-0.60
	MI(M=10)	3.92	1.91
$\bar{S}_{zx,N}$	FI(M=10)	3.51	1.22
	MI(M=10)	8.47	4.25
$\bar{Y}_{D,N}$	FI(M=10)	-6.12	-3.13
	MI(M=10)	14.68	7.29
$\bar{Z}_{D,N}$	FI(M=10)	-7.21	-3.97
	MI(M=10)	27.93	13.28
$\bar{P}_{y,N}$	FI(M=10)	-1.51	-0.71
	MI(M=10)	14.35	6.49
$\bar{P}_{z,N}$	FI(M=10)	1.60	0.78
	MI(M=10)	24.00	12.02