Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Fractional hot deck imputation for robust inference under item nonresponse in survey sampling

by Jae Kwang Kim and Shu Yang

Release date: December 19, 2014

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 211-230
Statistics Canada, Catalogue No. 12-001-X

Fractional hot deck imputation for robust inference under item nonresponse in survey sampling

Jae Kwang Kim and Shu Yang[1]

Abstract

Parametric fractional imputation (PFI), proposed by Kim (2011), is a tool for general purpose parameter estimation under missing data. We propose a fractional hot deck imputation (FHDI) which is more robust than PFI or multiple imputation. In the proposed method, the imputed values are chosen from the set of respondents and assigned proper fractional weights. The weights are then adjusted to meet certain calibration conditions, which makes the resulting FHDI estimator efficient. Two simulation studies are presented to compare the proposed method with existing methods.

Key Words: EM algorithm; Kullback-Leibler information; Missing at random (MAR); Multiple imputation.

1 Introduction

Imputation is a popular method of compensating for item non-response in sample surveys. Let y be the study variable subject to non-response and $\mathbf{x}$ be the vector of auxiliary variables fully observed. A model on the conditional distribution $f(y|\mathbf{x})$ is often used to generate imputed values for missing y_i. Such model-based imputation method is well developed in the literature. Multiple imputation of Rubin (1987) is a Bayesian approach of model-based imputation. Monte Carlo EM of Wei and Tanner (1990) can be treated as a frequentist's approach of model-based imputation. Kim (2011) proposed parametric fractional imputation to handle multivariate missing data.

However, the model-based imputation method that generates imputed values from $f(y|\mathbf{x})$ is not a hot deck imputation in the sense that artificial values are constructed after the imputation. A desirable property of hot deck imputation is that all imputed values are observed values. For example, imputed values for categorical variables will also be categorical with the same number of categories as observed for the respondents. For this reason, hot deck imputation is the most popular imputation method, especially in household surveys. Nearest neighbor imputation method is also a hot deck imputation. Chen and Shao (2001), Beaumont and Bocci (2009), Kim, Fuller and Bell (2011) investigated nearest neighbor imputation in the context of survey sampling. Durrant (2009), Haziza (2009) and Andridge and Little (2010) provided comprehensive overviews of the hot-deck imputation methods in survey sampling.

Fractional hot deck imputation was proposed by Kalton and Kish (1984) to achieve efficiency in hot deck imputation. Kim and Fuller (2004) and Fuller and Kim (2005) provided a rigorous treatment of fractional hot deck imputation and discussed variance estimation. However, their approach is only applicable when $\mathbf{x}$ is categorical. For continuous covariate case, predictive mean matching can be treated as a nearest neighbor imputation method using the predicted value obtained from $f(y|\mathbf{x})$ but its statistical properties are not fully addressed in the literature.

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, IA 50011. E-mail: jkim@iastate.edu; Shu Yang, Department of Statistics, Iowa State University, Ames, IA 50011.

In this paper, we propose a new fractional hot deck imputation (FHDI) method based on a parametric model of $f(y|\mathbf{x})$ that allows continuous covariates. The proposed method has several advantages over the existing methods. First, it is a hot deck imputation preserving the correlation structure between the items. Second, it is robust in that the resulting estimator is less sensitive against the failure of the assumed model $f(y|\mathbf{x})$. Third, it provides consistent variance estimators for various parameters without requiring the congeniality condition of Meng (1994). Multiple imputation, however, requires the congeniality condition for the validity of variance estimation. When the congeniality condition does not hold, multiple imputation often leads to conservative inference, which in turn reduces test powers. See Section 5.2 for more details.

The paper is organized as follows. Section 2 describes the basic setup. The proposed method is presented in Section 3. The robustness of FHDI is discussed in Section 4. Results from two simulation studies are presented in Section 5 before some concluding remarks are made in Section 6.

2 Basic setup

Consider a finite population of N elements identified by a set of indices $U = \{1, 2, \ldots, N\}$ with N known. Associated with each unit i in the population are study variables, $\mathbf{x}_i$ and y_i, with $\mathbf{x}_i$ always observed and y_i subject to non-response. Let A denote the set of indices for the elements in a sample selected by a probability sampling mechanism. We are interested in estimating η, defined as a (unique) solution to the population estimating equation $\sum_{i=1}^{N} U(\eta; \mathbf{x}_i, y_i) = 0$. For example, a population mean can be obtained by letting $U(\eta; \mathbf{x}_i, y_i) = \eta - y_i$. Under complete response, a consistent estimator of η is obtained by solving

$$\sum_{i \in A} w_i U(\eta; \mathbf{x}_i, y_i) = 0, \tag{2.1}$$

where $w_i = \{Pr(i \in A)\}^{-1}$ is the inverse of the first-order inclusion probability of unit i. Binder and Patak (1994) and Rao, Yung and Hidiroglou (2002) considered the asymptotic properties of the estimator obtained from (2.1). Under the existence of missing data, we define

$$\delta_i = \begin{cases} 1 & \text{if } y_i \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

A consistent estimator of η is then obtained by taking the conditional expectation and solving

$$\sum_{i \in A} w_i \left[\delta_i U(\eta; \mathbf{x}_i, y_i) + (1 - \delta_i) E\{U(\eta; \mathbf{x}_i, Y) | \mathbf{x}_i, \delta_i = 0\} \right] = 0 \tag{2.2}$$

for η. Estimating equation (2.2) is sometimes referred to as expected estimating equation (Wang and Pepe 2000).

To compute the conditional expectation in (2.2), we assume that the finite population at hand is a realization from an infinite population, called superpopulation. In the superpopulation model, we often postulate a parametric conditional distribution of y given $\mathbf{x}$, $f(y|\mathbf{x}; \theta)$, which is known up to the parameter θ with parameter space Ω. Under the specified model, we can compute a consistent estimator

$\hat{\theta}$ of θ and then use a Monte Carlo method to evaluate the conditional expectation in (2.2) given the estimate $\hat{\theta}$. If the response mechanism is missing at random (MAR) or ignorable in the sense of Rubin (1976), we can approximate the expected estimating equation in (2.2) by

$$\sum_{i\in A} w_i \left\{ \delta_i U\left(\eta; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \frac{1}{m} \sum_{j=1}^{m} U\left(\eta; \mathbf{x}_i, y_i^{*(j)}\right) \right\} = 0, \tag{2.3}$$

where

$$y_i^{*(1)}, \ldots, y_i^{*(m)} \overset{i.i.d.}{\sim} f\left(y_i \mid \mathbf{x}_i; \hat{\theta}\right).$$

Often, we use the maximum likelihood estimator $\hat{\theta}$, which solves

$$S\left(\theta\right) = \sum_{i\in A} w_i \delta_i S\left(\theta; \mathbf{x}_i, y_i\right) = 0, \tag{2.4}$$

where $S\left(\theta; \mathbf{x}, y\right) = \partial \log f\left(y \mid \mathbf{x}; \theta\right) / \partial\theta$. Note that we use the sampling weights w_i in the score equation (2.4). Thus, we are implicitly assuming that the imputation model, the model for generating the imputed values, is the model about the finite population values $f\left(y_i \mid \mathbf{x}_i\right)$, not the model about the sample values. Thus, we allow that the sampling mechanism can be informative in the sense of Pfeffermann (2011). Multiple imputation, on the other hand, uses the sample model, $f_s\left(y_i \mid \mathbf{x}_i\right) \equiv f\left(y_i \mid \mathbf{x}_i, i \in A\right)$, to generate the imputed values and often assumes that the sampling mechanism is non-informative. Thus, in multiple imputation, MAR is assumed for the sample at hand, while, in fractional imputation, MAR is assumed for the population. Under informative sampling design, generating imputed values from the sample model $f_s\left(y_i \mid \mathbf{x}\right)$ does not necessarily lead to valid inference even when sample MAR condition holds. See Section 8.4 of Kim and Shao (2013) for further discussion of MAR under informative sampling.

To compute the conditional expectation in (2.2) efficiently, the parametric fractional imputation (PFI) of Kim (2011) can be used. In PFI, the imputed values are generated from a suitable proposal distribution $h\left(y \mid \mathbf{x}_i\right)$ and then the imputed estimating equation (2.3) is changed to

$$\sum_{i\in A} w_i \left\{ \delta_i U\left(\eta; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \sum_{j=1}^{m} w_{ij}^{*} U\left(\eta; \mathbf{x}_i, y_i^{*(j)}\right) \right\} = 0, \tag{2.5}$$

where

$$w_{ij}^{*} = \frac{f\left(y_i^{*(j)} \mid \mathbf{x}_i; \hat{\theta}\right) \Big/ h\left(y_i^{*(j)} \mid \mathbf{x}_i\right)}{\sum_{k=1}^{m} \left\{ f\left(y_i^{*(k)} \mid \mathbf{x}_i; \hat{\theta}\right) \Big/ h\left(y_i^{*(k)} \mid \mathbf{x}_i\right) \right\}}. \tag{2.6}$$

The choice of the proposal distribution $h\left(\cdot\right)$ is somewhat arbitrary. We will discuss a particular choice that may lead to a robust estimation.

The consistency of the resulting estimator $\hat{\eta}$ from (2.3) or (2.5) can be established under the assumption that the conditional distribution $f\left(y \mid \mathbf{x}; \theta\right)$ is correctly specified (by similar argument in the proof of Corollary II.2 of Andersen and Gill (1982) and its proof is skipped here). In this paper, we consider an alternative approach of fractional imputation that is more robust against the failure of the assumption on the imputation model.

3 Proposed method

We first consider a particular fractional hot deck imputation method, called **full fractional imputation**, where the imputed values are taken from the set of respondents denoted as $A_R = \{i \in A; \delta_i = 1\}$. That is, the j-th imputed value of missing y_i, denoted by $y_i^{*(j)}$, is equal to the j-th value of y among the set in A_R. We propose a fractional hot deck imputation approach that makes use of the parametric model assumption $f(y \mid \mathbf{x}; \theta)$. If all of the elements in A_R are selected as the imputed values for missing y_i, we can treat $\{y_j; j \in A_R\}$ as a realization from $f(y_j \mid \delta_j = 1)$ and fractional weight assigned to donor y_j for the missing item y_i is, by choosing $h(y_j \mid \mathbf{x}_i) = f(y_j \mid \delta_j = 1)$ in (2.6),

$$\begin{aligned} w_{ij}^* &\propto f(y_j \mid \mathbf{x}_i, \delta_i = 0; \hat{\theta}) \Big/ f(y_j \mid \delta_j = 1) \\ &\propto f(y_j \mid \mathbf{x}_i; \hat{\theta}) \Big/ f(y_j \mid \delta_j = 1), \end{aligned} \tag{3.1}$$

with $\sum_{j;\delta_j=1} w_{ij}^* = 1$, and $\hat{\theta}$ being the MLE obtained from (2.4). The second line follows from the MAR assumption. Furthermore, we can write

$$\begin{aligned} f(y_j \mid \delta_j = 1) &= \int f(y_j \mid \mathbf{x}, \delta_j = 1) f(\mathbf{x} \mid \delta_j = 1) d\mathbf{x} \\ &= \int f(y_j \mid \mathbf{x}) f(\mathbf{x} \mid \delta_j = 1) d\mathbf{x} \\ &\cong \frac{1}{N_R} \sum_{k=1}^{N} \delta_k f(y_j \mid \mathbf{x}_k), \end{aligned} \tag{3.2}$$

where the second equality follows from the MAR assumption, and the last (approximate) equality follows by approximating the integral by the population empirical distribution, and N_R is the number of respondents in the population. Using the survey weights, we can approximate

$$f(y_j \mid \delta_j = 1) \cong \frac{\sum_{k \in A_R} w_k f(y_j \mid \mathbf{x}_k)}{\sum_{k \in A_R} w_k}$$

and the fractional weights in (3.1) are computed from

$$w_{ij}^* \propto \frac{f(y_j \mid \mathbf{x}_i; \hat{\theta})}{\sum_{k \in A_R} w_k f(y_j \mid \mathbf{x}_k; \hat{\theta})} \tag{3.3}$$

with $\sum_{j \in A_R} w_{ij}^* = 1$. In (3.3), the point mass w_{ij}^* assigned to donor y_j for missing unit i is expressed by the ratio of the density $f(y \mid \mathbf{x})$. Thus, for each missing unit i, $n_R = |A_R|$ observations are used as donors for the hot deck imputation using w_{ij}^* as the fractional weights. Such fractional imputation can be called full fractional imputation (FFI) because there is no randomness due to the imputation mechanism. The FFI estimator of η, defined by $\sum_{i=1}^{N} U(\eta; \mathbf{x}_i, y_i) = 0$, is then computed by solving

$$\sum_{i \in A} w_i \left\{ \delta_i U(\eta; \mathbf{x}_i, y_i) + (1 - \delta_i) \sum_{j \in A_R} w_{ij}^* U(\eta; \mathbf{x}_i, y_j) \right\} = 0, \tag{3.4}$$

where w_{ij}^* is defined in (3.3). Note that the imputed estimating equation (3.4) is a good approximation to the expected estimating equation in (2.2).

In survey sampling, an imputed data set with a large imputation size may not be desirable. Thus, instead of taking all the observations in A_R as donors for each missing item, a subset of A_R can be selected to reduce the size of the donor set of missing y_i. Thus, the selection of the donors is viewed as a sampling problem and we use an efficient sampling design and weighting techniques to obtain efficient imputation estimators. For the donor selection mechanism, efficient sampling designs, such as a stratified sampling design or systematic Proportional-to-Size (PPS) sampling, can be used to select donors of size m. A systematic PPS sampling for fractional hot deck imputation can be described as follows:

1. Within each i with $\delta_i = 0$, sort the donors in the full respondent set $\{y_j; \delta_j = 1\}$ in ascending order as $y_{(1)} \le \cdots \le y_{(r)}$ and use $w_{i(j)}^*$ to denote the fractional weight associated with $y_{(j)}$. That is, $w_{i(j)}^* = w_{ik}^*$ for $y_{(j)} = y_k$.

2. Partition $[0,1]$ by $\left\{I_j \equiv \left[\sum_{k=0}^{j} w_{i(j)}^*, \sum_{k=0}^{j+1} w_{i(j)}^*\right), j = 1,\ldots,r-1\right\}$, where $w_{i(0)}^* = 0$.

3. Generate $u \sim \text{uniform}(0,1/m)$ and let $u_k = u + k/m$, $k = 0,\ldots,m-1$. For $k = 0,\ldots,m-1$, if $u_k \in I_j$ for some $0 \le j \le r-1$, include j in the sample D_i.

After we select D_i from the complete set of respondents, the selected donors in D_i are assigned with the initial fractional weights $w_{ij0}^* = 1/m$. The fractional weights are further adjusted to satisfy

$$\sum_{i\in A} w_i \left\{(1-\delta_i)\sum_{j\in D_i} w_{ij,c}^* \mathbf{q}(\mathbf{x}_i, y_j)\right\} = \sum_{i\in A} w_i \left\{(1-\delta_i)\sum_{j\in A_R} w_{ij}^* \mathbf{q}(\mathbf{x}_i, y_j)\right\}, \tag{3.5}$$

for some $\mathbf{q}(\mathbf{x}_i, y_j)$, and $\sum_{j\in D_i} w_{ij,c}^* = 1$ for all i with $\delta_i = 0$, where w_{ij}^* is the fractional weights for FFI method, as defined in (3.3). Regarding the choice of the control function $\mathbf{q}(\mathbf{x}, y)$ in (3.5), we can use $\mathbf{q}(\mathbf{x}, y) = (y, y^2)'$, which keeps the empirical distributions of y for D_i and A_R as close as possible in the sense that the first and second moment of y are the same. Other choices can also be considered. See Fuller and Kim (2005).

The problem of adjusting the initial weights to satisfy certain constraints is often called calibration and the resulting fractional weights can be called calibrated fractional weights. Using the idea of regression weighting, the final calibration fractional weights that satisfy (3.5) and $\sum_j w_{ij,c}^* = 1$ can be computed by

$$w_{ij,c}^* = w_{ij0}^* + w_{ij0}^* \Delta\left(\mathbf{q}_{ij}^* - \bar{\mathbf{q}}_{i\cdot}^*\right), \tag{3.6}$$

where $\mathbf{q}_{ij}^* = \mathbf{q}(\mathbf{x}_i, y_j)$, $\bar{\mathbf{q}}_{i\cdot}^* = \sum_{j\in A_R} w_{ij0}^* \mathbf{q}_{ij}^*$,

$$\Delta = \left\{C_q - \sum_{i\in A} w_i (1-\delta_i) \sum_{j\in A_R} w_{ij0}^* \mathbf{q}_{ij}^*\right\}^T \left\{\sum_{i\in A} w_i (1-\delta_i) \sum_{j\in A_R} w_{ij0}^* \left(\mathbf{q}_{ij}^* - \bar{\mathbf{q}}_{i\cdot}^*\right)^{\otimes 2}\right\}^{-1}$$

and $C_q = \sum_{i\in A} w_i \left\{(1-\delta_i)\sum_{j\in A_R} w_{ij}^* \mathbf{q}(\mathbf{x}_i, y_j)\right\}$. Here, $B^{\otimes 2}$ denotes BB^T. Some of the fractional weights computed by (3.6) can take negative values. If that happens, algorithms alternative to regression weighting should be used. For example, consider entropy weighting, where the fractional weights of the form

$$w_{ij,c}^* = \frac{w_{ij}^* \exp\left(\Delta \mathbf{q}_{ij}^*\right)}{\sum_{k\in A_R} w_{ik}^* \exp\left(\Delta \mathbf{q}_{ik}^*\right)} \tag{3.7}$$

are approximately equal to the regression fractional weights in (3.6) and are always positive. Once the calibration fractional weights are obtained, the FHDI estimator of η is then computed by solving

$$\sum_{i\in A} w_i \left\{\delta_i U\left(\eta; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right)\sum_{j\in D_i} w_{ij,c}^* U\left(\eta; \mathbf{x}_i, y_j\right)\right\} = 0. \tag{3.8}$$

For variance estimation, a replication method can be used. See Appendix A.1 for a brief discussion of the replication variance estimator for the proposed method.

Furthermore, the proposed method can handle non-ignorable non-response under the correct specification of the response model. See Appendix A.3 for the extension to non-ignorable non-response case.

4 Robustness

We now discuss the robustness of the proposed method against a small departure from the assumed parametric model. The robustness feature in our proposed estimator is defined to be robust against imputation model misspecification, a small exponential tilting of the true model. For simplicity of the presentation, assume that the sampling design is simple random sampling and the realized sample is a random sample from the superpopulation model.

We assume that the true model $g(y\,|\,x)$ does not belong to $\left\{f(y\,|\,x;\theta); \theta\in\Omega\right\}$. However, we can still specify a working model $f(y\,|\,x;\theta)$ and compute the MLE of θ. It is well known (White 1982) that the MLE converges to θ^*, the minimizer of the Kullback-Leibler information

$$K(\theta) = E_g\left[\log\left\{\frac{g(Y\,|\,x)}{f(Y\,|\,x;\theta)}\right\}\right]$$

for $\theta\in\Omega$. Sung and Geyer (2007) discussed the asymptotic properties of the Monte Carlo MLE of θ under missing data.

To formally discuss robustness, suppose that the true distribution $g(y\,|\,x)$ belongs to the neighborhood

$$\mathcal{N}_\varepsilon = \left\{g; D(g, f) < \frac{1}{2}\varepsilon^2\right\} \tag{4.1}$$

for some radius $\varepsilon > 0$, where

$$D(g, f) = \int log\left(\frac{g}{f}\right) g\ dy, \tag{4.2}$$

is the Kullback-Leibler distance measure. The neighborhood (4.1) can be characterized in the following way. Let $z(x,y,\theta)$ be a function of x, y and θ, standardized to satisfy $E_{Y|x}(z)=0$ and $Var_{Y|x}(z)=1$, and define

$$g(y\mid x)= f(y\mid x;\theta)\exp\{\varepsilon z(x,y,\theta)-\kappa(x,\theta)\}, \tag{4.3}$$

where

$$\kappa=\log\Big(E_{Y|x}\big[\exp\{\varepsilon z(x,Y,\theta)\}\big]\Big).$$

For small $\varepsilon>0$ it can be shown that

$$\kappa\cong D(g,f)\cong\frac{1}{2}\varepsilon^2. \tag{4.4}$$

Equation (4.3) represents an extensive set of distributions close to $f(y\mid x;\theta)$ created by varying $z(x,y,\theta)$ over different standardized functions, where z and ε contain some geometric interpretation which represent the direction and magnitude of the misspecification respectively. For p-dimension parameter θ, we can specify the directions of the misspecification as

$$\big(z_1,z_2,\ldots,z_p\big)^T=I_\theta^{-1/2}s(x,y,\theta),$$

where $s(x,y,\theta)=\partial\log f(y\mid x;\theta)/\partial\theta$ and I_θ is the information matrix for θ. Represent $z(x,y,\theta)$ as

$$z(x,y,\theta)=\lambda^T I_\theta^{-1/2}s(x,y,\theta),$$

where $\sum_{i=1}^p\lambda_i^2=1$, then $z(x,y,\theta)$ satisfies the standardization criterion of $E_{Y|x}(z)=0$ and $Var_{Y|x}(z)=1$. See Copas and Eguchi (2001) for further discussion of this expression.

Let $w_{ij,g}^*$ be the fractional weight of the form (3.3) using the true density g and $w_{ij,f}^*$ be the corresponding fractional weight using the "working density" f. By the special construction of the weights, we can establish

$$w_{ij,g}^*\cong w_{ij,f}^*+\varepsilon\lambda^T I_\theta^{-1/2}\frac{\partial}{\partial\theta}\big(w_{ij,f}^*\big). \tag{4.5}$$

Proof of (4.5) is given in Appendix A.2. Thus

$$\begin{aligned}\sum_i w_i\sum_j w_{ij,g}^*U(\eta;x_i,y_j)\cong\ & \sum_i w_i\sum_j w_{ij,f}^*U(\eta;x_i,y_j)\\ &+\varepsilon\lambda^T I_\theta^{-1/2}\sum_i w_i\sum_j\frac{\partial}{\partial\theta}\big(w_{ij,f}^*\big)U(\eta;x_i,y_j).\end{aligned} \tag{4.6}$$

For small ε, we have

$$\sum_i w_i\sum_j w_{ij,g}^*U(\eta;x_i,y_j)\cong\sum_i w_i\sum_j w_{ij,f}^*U(\eta;x_i,y_j),$$

and so the resulting estimator of η from $\sum_i w_i\sum_j w_{ij,f}^*U(\eta;x_i,y_j)=0$ will be close to the true value η_0.

5 Simulation study

We performed two simulation studies. In Section 5.1, we compared the performance of the proposed method with some other imputation methods in a correctly specified model and a misspecified model, respectively, with ignorable missing data. In Section 5.2, we compared the statistical power of a test based on FHDI versus MI.

5.1 Simulation one

The first simulation study tested the performance of the proposed method under the setup of ignorable missing data. We used two sets of models to generate the observations. In model A, $y_i = 0.5x_i + e_i$, where $x_i \sim \exp(1)$, $e_i \sim N(0,1)$, with x_i and e_i being independent. In model B, $y_i = 0.5x_i + e_i$, where $x_i \sim \exp(1)$, $e_i \sim \{\chi^2(2) - 2\}/2$, with x_i and e_i being independent. Random samples of size $n = 200$ were separately generated from the two models. In addition to (x_i, y_i), we also generated δ_i from Bernoulli(π_i), where $\pi_i = \{1 + \exp(-0.2 - x_i)\}^{-1}$. Variable x_i was always observed but variable y_i was observed if and only if $\delta_i = 1$. The overall response rates were about 65% in both cases. We used $B = 2,000$ Monte Carlo samples in the simulation.

From each of the Monte Carlo samples, one generated from model A and the other generated from model B, we computed the following eight estimators:

1. Full sample estimator (Full) that is computed using the full sample.

2. Predictive Mean Matching (PMM) is a semi-parametric imputation method, which fills in a value randomly from observations that are closest to the predicted value obtained from $f(y \mid \mathbf{x})$. The PMM was implemented using "mice.impute.pmm" function in R.

3. Multiple imputation (MI) estimator with imputation size $m = 10$, where the imputed values are generated from the normal-theory regression model, as considered in Schenker and Welsh (1988).

4. Parametric fractional imputation (PFI) estimator without calibration with imputation size $m = 10$.

5. Parametric fractional imputation (PFI_cal) estimator with calibration with imputation size $m = 10$. The fractional weights are computed using the calibration method in (3.6) with $\mathbf{q} = (y, y^2)$.

6. Full fractional imputation (FFI) estimator using the full set of respondents as imputation values, i.e. the imputation size $m = n_R$, where n_R is the size of A_R.

7. Fractional hot deck imputation (FHDI) estimator without calibration using a small subset of respondents of size $m = 10$ as imputation values.

8. Fractional hot deck imputation (FHDI_cal) estimator with calibration using a small subset of respondents of size $m = 10$ as imputation values. The fractional weights are computed using the calibration method in (3.6) with $\mathbf{q} = \left(y, y^2\right)$.

Multiple imputation is an approach of generating imputed values with simplified variance estimation. In this procedure, Bayesian methods of generating imputed values are considered, where $m > 1$ imputed values are generated from the posterior predictive distribution. Using the imputed values $\mathbf{y}^{*(1)},\ldots,\mathbf{y}^{*(m)}$, the multiple imputation estimator of η, denoted by $\hat{\eta}_{MI}$ is

$$\hat{\eta}_{MI} = \frac{1}{m}\sum_{k=1}^{m}\hat{\eta}^{(k)}$$

where $\hat{\eta}^{(k)}$ is the complete response estimator applied to the k-th imputed data set. Rubin's formula can be used for variance estimation in MI,

$$\hat{V}_{MI}\left(\hat{\eta}_{MI}\right) = W_m + \left(1+\frac{1}{m}\right)B_m, \tag{5.1}$$

where $W_m = m^{-1}\sum_{k=1}^{m}\hat{V}^{(k)}$, $B_m = \left(m-1\right)^{-1}\sum_{k=1}^{m}\left(\hat{\eta}^{(k)} - \hat{\eta}_{MI}\right)^2$, and $\hat{V}^{(k)}$ is the variance estimator of $\hat{\eta}^{(k)}$ under complete response applied to the k-th imputed data set.

In both models, we used the normal density with mean $\beta_0 + \beta_1 x$ and variance σ^2 as the working model for imputation. Thus, the working model is the true model in model A but not true in model B.

We considered three parameters: $\theta_1 = E\left(Y\right)$, the population mean of y, $\theta_2 = Pr\left(Y < 1\right)$, the proportion of Y less than one, and θ_3, the 0.5 quantile of Y. In estimating θ_2 under full sample, we used $\hat{\theta}_{2,n} = n^{-1}\sum_{i=1}^{n} I\left(y_i < 1\right)$. In estimating θ_3 under full sample, we used $\hat{\theta}_{3,n} = \hat{F}^{-1}\left(p\right) = \inf\left\{y : \hat{F}\left(y\right) > p\right\}$, where $\hat{F}\left(y\right) = n^{-1}\sum_{i=1}^{n} I\left(y_i < y\right)$ and $p = 0.5$.

Table 5.1 and Table 5.2 show Monte Carlo means, standardized variance (Std Var) and standardized mean squared errors (Std MSE) of the eight estimators under model A and under model B, respectively. The standardized variance (mean squared error) is calculated as the ratio of variance (mean squared error) and the variance (mean squared error) of the full sample estimator multiplied by 100, which measures the increased variance (mean squared error) due to imputation relative to the full sample estimator. As for the Monte Carlo means (4th column), the imputation estimators are all unbiased for estimating θ_1, θ_2, and θ_3 under model A. Under model B, PMM, MI, PFI, PFI_cal for estimating θ_3 have much larger biases in absolute values than FFI, FHDI, and FHDI_cal under model misspecification in this simulation. Regarding the standardized variance and standardized mean squared error (5th and 6th column), PFI is more efficient than FHDI. The reason is that in PFI, the imputed values are generated according to the conditional distribution $f\left(y \mid x\right)$ directly; whereas in FHDI, the imputed values can be taken from respondents with dominantly large fractional weights. The effective imputation data size is determined by the imputed observations with large fractional weights, which also contribute to the loss of efficiency. FHDI loses efficiency in order to gain robustness. Lastly, FHDI with $m = 10$ has slightly larger standardized variance for θ_2 than FFI, because of the additional variability due to the sampling procedure. Comparing PFI with PFI_cal and FHDI with FHDI_cal, the calibration step improves the efficiency a little bit. The PMM shows the largest variance in all scenarios.

Table 5.1
Monte Carlo mean, standardized variance and standardized mean squared error of point estimators in Model A of Simulation one.

Model	Parameter	Method	Mean	Std Var	Std MSE
A	μ_y	Full	0.50	100	100
		PMM	0.50	175	175
		MI ($m = 10$)	0.50	135	135
		PFI ($m = 10$)	0.50	130	130
		PFI cal ($m = 10$)	0.50	130	130
		FFI ($m = n_R$)	0.50	130	130
		FHDI ($m = 10$)	0.50	156	156
		FHDI cal ($m = 10$)	0.50	130	130
	$Pr(Y < 1)$	Full	0.68	100	100
		PMM	0.68	168	167
		MI ($m = 10$)	0.68	112	112
		PFI ($m = 10$)	0.68	110	110
		PFI cal ($m = 10$)	0.68	109	109
		FFI ($m = n_R$)	0.68	130	130
		FHDI ($m = 10$)	0.68	137	136
		FHDI cal ($m = 10$)	0.68	132	132
	Quantile	Full	0.47	100	100
		PMM	0.47	184	184
		MI ($m = 10$)	0.47	111	111
		PFI ($m = 10$)	0.47	111	111
		PFI cal ($m = 10$)	0.47	111	111
		FFI ($m = n_R$)	0.47	135	135
		FHDI ($m = 10$)	0.47	142	142
		FHDI cal ($m = 10$)	0.47	141	141

Table 5.2
Monte Carlo mean, standardized variance and standardized mean squared error of point estimators in Model B of Simulation one.

Model	Parameter	Method	Mean	Std Var	Std MSE
B	μ_y	Full	0.50	100	100
		PMM	0.50	172	172
		MI ($m = 10$)	0.50	131	131
		PFI ($m = 10$)	0.50	131	131
		PFI cal ($m = 10$)	0.50	128	128
		FFI ($m = n_R$)	0.50	127	127
		FHDI ($m = 10$)	0.50	147	147
		FHDI cal ($m = 10$)	0.50	127	127
	$Pr(Y < 1)$	Full	0.75	100	100
		PMM	0.75	166	166
		MI ($m = 10$)	0.73	140	170
		PFI ($m = 10$)	0.73	138	168
		PFI cal ($m = 10$)	0.73	137	169
		FFI ($m = n_R$)	0.75	137	137
		FHDI ($m = 10$)	0.75	145	145
		FHDI cal ($m = 10$)	0.75	140	141
	Quantile	Full	0.26	100	100
		PMM	0.24	191	198
		MI ($m = 10$)	0.31	122	159
		PFI ($m = 10$)	0.31	123	160
		PFI cal ($m = 10$)	0.31	122	159
		FFI ($m = n_R$)	0.26	135	135
		FHDI ($m = 10$)	0.26	144	144
		FHDI cal ($m = 10$)	0.26	139	139

For variance estimation, we considered replication variance estimation for FFI and FHDI, particularly the delete-1 Jackknife variance estimation, which is described in Appendix A.1. We also considered variance estimation in MI, which uses Rubin's formula (5.1).

Table 5.3 shows the Monte Carlo relative biases of the variance estimators, which is calculated as $\left[E_{MC}\left\{\hat{V}\right\}-V_{MC}\left\{\hat{\theta}\right\}\right]\Big/V_{MC}\left\{\hat{\theta}\right\}$, where $E_{MC}\left\{\hat{V}\right\}$ is the Monte Carlo mean of variance estimates $\hat{V}$, and $V_{MC}\left\{\hat{\theta}\right\}$ is the Monte Carlo variance of the point estimates $\hat{\theta}$. The relative bias of the variance estimator in FFI and FHDI is reasonably small for all parameters considered in both models, suggesting that the replication variance estimator is valid. The relative bias and $t-$statistics of variance estimator in MI are small for θ_1 but quite large for θ_2 even when the working model is true (model A). Rubin's formula is based on the following decomposition,

$$V\left(\hat{\theta}_{MI}\right)=V\left(\hat{\theta}_n\right)+V\left(\hat{\theta}_{MI}-\hat{\theta}_n\right), \tag{5.2}$$

where $\hat{\theta}_n$ is the full sample estimator of η. Basically, the W_m term in (5.1) estimates $V\left(\hat{\theta}_n\right)$ and the $\left(1+m^{-1}\right)B_m$ term in (5.1) estimates $V\left(\hat{\theta}_{MI}-\hat{\theta}_n\right)$. The decomposition (5.2) holds when $\hat{\theta}_n$ is the MLE of θ, which is the congeniality condition of $\hat{\theta}_n$ (Meng 1994). For general case, we have

$$V\left(\hat{\theta}_{MI}\right)=V\left(\hat{\theta}_n\right)+V\left(\hat{\theta}_{MI}-\hat{\theta}_n\right)+2Cov\left(\hat{\theta}_{MI}-\hat{\theta}_n,\hat{\theta}_n\right) \tag{5.3}$$

and Rubin's variance estimator can be biased if $Cov\left(\hat{\theta}_{MI}-\hat{\theta}_n,\hat{\theta}_n\right)\neq 0$. The congeniality condition holds true for estimating the population mean; however, it does not hold for the method of moments estimator of $Pr\left(Y<1\right)$. Note that the imputed estimator of $\theta_2=Pr\left(Y<1\right)$ can be expressed as

$$\hat{\theta}_{2,I}=n^{-1}\sum_{i=1}^{n}\left[\delta_i I\left(y_i<1\right)+\left(1-\delta_i\right)E\left\{I\left(y_i<1\right)\mid x_i;\hat{\mu},\hat{\sigma}\right\}\right]. \tag{5.4}$$

Thus, the imputed estimators of θ_2 "borrows strength" by making use of extra information associated with $f\left(y\mid x\right)$. That is, the normality of $f\left(y\mid x\right)$ is used in computing the conditional expectation in (5.4), which improves the efficiency of the imputed estimator for θ_2. The same phenomenon also holds for θ_3. In Table 5.1, the increase of variance due to imputation for MI with $m=10$ is about 35 % for θ_1 but only 12% and 11% for θ_2 and θ_3, respectively, which shows the phenomenon of "borrowing strength" for estimating θ_2 and θ_3 thanks to the use of extra information in the imputation stage. Thus, when the congeniality conditions do not hold, the imputed estimator improves the efficiency but Rubin's variance estimator does not recognize this improvement.

Table 5.3
Monte Carlo relative bias of the replication variance estimator in Simulation one.

Model	Parameter	Method	R.B. (%)
*A	$V(\hat{\theta}_1)$	MI ($m=10$)	-2.33
		FFI ($m=n_R$)	-0.80
		FHDI_cal ($m=10$)	-0.80
	$V(\hat{\theta}_2)$	MI ($m=10$)	8.20
		FFI ($m=n_R$)	-5.01
		FHDI_cal ($m=10$)	-5.12
	$V(\hat{\theta}_3)$	MI ($m=10$)	19.84
		FFI ($m=n_R$)	4.50
		FHDI_cal ($m=10$)	3.78
*B	$V(\hat{\theta}_1)$	MI ($m=10$)	2.60
		FFI ($m=n_R$)	-0.56
		FHDI_cal ($m=10$)	-0.56
	$V(\hat{\theta}_2)$	MI ($m=10$)	-3.33
		FFI ($m=n_R$)	-1.89
		FHDI_cal ($m=10$)	-3.25
	$V(\hat{\theta}_3)$	MI ($m=10$)	-8.99
		FFI ($m=n_R$)	3.50
		FHDI_cal ($m=10$)	3.80

5.2 Simulation two

Simulation two tested the power of the proposed method in a hypothesis test using the null model as the imputation model. Samples of bivariate data (x_i, y_i) of size $n=100$ were generated from

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 \left(x_i^2 - 1\right) + e_i \tag{5.5}$$

where $(\beta_0, \beta_1, \beta_2) = (0, 0.9, 0.06)$, $x_i \sim N(0,1)$, $e_i \sim N(0, 0.16)$, with x_i and e_i being independent. The variable x_i is always observed but the probability that y_i responds is 0.5. Monte Carlo samples were generated independently for $B=10,000$ times. We are interested in testing $H_0: \beta_2 = 0$ from the respondents. We compared FHDI with MI using the same imputation size $m=30$. The imputation model is the null model,

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

That is, the imputation model uses extra information of $\beta_2 = 0$. From the imputed data, we fit model (5.5) and computed the power of a test $H_0: \beta_2 = 0$ at the significant level of 0.05. In addition, we also considered the complete case (CC) method that only uses the respondents for regression.

Table 5.4 shows the Monte Carlo mean and variance of the point estimators, relative bias of the variance estimator and the Monte Carlo power of testing $H_0: \beta_2 = 0$. In each Monte Carlo sample, we constructed a 95% Wald confidence interval of β_2 as $\left(\hat{\beta}_2 - 1.96\hat{V}^{1/2}, \hat{\beta}_2 + 1.96\hat{V}^{1/2}\right)$ and reject the null hypothesis if $\beta_2 = 0$ does not fall in the Wald confidence interval. The Monte Carlo power is calculated as

the relative frequency of rejecting the null hypothesis among the Monte Carlo samples. From the second column, FHDI and MI estimators are biased for β_2, as expected since in imputation the imputation model is the null model and it is slightly different from the true model that generated sample. The bias of FHDI is smaller than that of MI because of the robustness of FHDI discussed in Section 4. In MI, 50% of the imputed MI data comes from the null model and the other 50% from the true model, so the slope β_2 is attenuated to zero by half of the true slope. In FHDI, though we used the null model to calculate the fractional weights, the imputed data come from the true model which reduces the bias. Moreover, MI provides more efficient point estimators than the CC method but variance estimation is very conservative (about 180% 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. On the other hand, FHDI also provides more efficient point estimators than the CC method and variance estimation is essentially unbiased, the statistical power of the test based on FHDI is higher than the CC method.

Table 5.4
Simulation results based on 10,000 Monte Carlo samples in Simulation two.

Method	$E(\hat{\beta}_2)$	$V(\hat{\beta}_2)$	R.B.$(\hat{V})$	Power
FHDI	0.046	0.00146	0.02	0.314
MI	0.028	0.00056	1.81	0.044
CC	0.060	0.00234	-0.01	0.285

6 Concluding remarks

We have proposed a fractional hot deck imputation method that uses a parametric model for $f(y|\mathbf{x})$ when $\mathbf{x}$ contains continuous components. The proposed method provides robust estimation for the parameters in the sense that the imputation model is not necessarily equal to the data-generating model. The price we pay in the FHDI is the loss of efficiency in point estimation. Under our first simulation, the FHDI estimator for $P(Y<1)$ has the second largest variance but the smallest mean squared error when the working model is not true, as compared with other estimators.

The loss of efficiency mainly comes from the fact that the fractional weights are more variable than those under the PFI method because some of $\mathbf{x}_j$ are not useful in imputing y_i. That is, the value of $f(y_i|\mathbf{x}_j;\hat{\theta})$ can be very small. The fractional hot deck imputation under a small imputation size (e.g. $m=10$) does not increase the variance significantly, as can be seen in Table 5.1 under model A.

The proposed fractional imputation method can actually be used to develop a single imputation method by applying FHDI with $m=1$, which selects an imputed value with probability proportional to the fractional weight for each missing unit. In this case, the FHDI can be used to develop a single imputation that is still robust against model misspecification. However, weighting calibration cannot co-exist with single imputation. Calibration constraints can still be achieved by employing the balanced imputation method as discussed in Chauvet, Deville and Haziza (2011) or the rejective Poisson sampling of Fuller (2009). Further investigation along this direction will be a topic of future research.

Acknowledgements

We thank two anonymous referees and the associate editor for very helpful comments. This research was partially supported by a grant from NSF (MMS-121339) and by the Cooperative Agreement between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University.

Appendix

A.1 Replication variance estimation

For variance estimation, replication methods can be used. Let $w_i^{[k]}$ be the k-th replication weights such that

$$\hat{V}_{rep} = \sum_{k=1}^{L} c_k \left(\hat{Y}^{[k]} - \hat{Y}\right)^2$$

is consistent for the variance of $\hat{Y} = \sum_{i \in A} w_i y_i$, where L is the replication size, c_k is the k-th replication factor that depends on the replication method and the sampling mechanism, and $\hat{Y}^{[k]} = \sum_{i \in A} w_i^{[k]} y_i$ is the k-th replicate of $\hat{Y}$. In delete-1 jackknife variance estimation, $L = n$ and $c_k = (n-1)/n$.

To apply the replication method in FFI, we first apply the replication weights $w_i^{[k]}$ in (2.4) to compute $\hat{\theta}^{[k]}$. Once $\hat{\theta}^{[k]}$ is obtained, we use the same imputed values to compute the initial replication fractional weights

$$w_{ij}^{*[k]} \propto w_j^{[k]} w_j^{-1} f\left(y_j \mid x_i; \hat{\theta}^{[k]}\right) \Bigg/ \left\{ \sum_{l \in A_R} w_l^{[k]} f\left(y_j \mid x_l; \hat{\theta}^{[k]}\right) \right\}, \tag{A.1}$$

with $\sum_{j \in A_R} w_{ij}^{*[k]} = 1$. The variance of $\hat{\eta}_{FFI}$, computed from (3.4), is then computed by

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

where $\hat{\eta}_{FFI}^{[k]}$ comes from solving

$$\sum_{i \in A} w_i^{[k]} \left\{ \delta_i U\left(\eta; \mathbf{x}_i, y_i\right) + \left(1 - \delta_i\right) \sum_{j \in A_R} w_{ij}^{*[k]} U\left(\eta; \mathbf{x}_i, y_j\right) \right\} = 0,$$

and $w_{ij}^{*[k]}$ is defined in (A.1).

We now discuss replication variance estimation of the FHDI estimator $\hat{\eta}_{FHDI}$ computed from (3.8). Define $d_{ij} = 1$ if $j \in D_i$ and $d_{ij} = 0$ otherwise. Note that $\hat{\eta}_{FHDI}$ is computed via two steps: in the first step, a systematic PPS sampling is used with the selection probability proportional to the fractional weights from the FFI method. In the second step, the calibration weighting method using the constraint (3.5) with

$\sum_{j\in A_R} d_{ij} w_{ij,c}^{*} = 1$ is used. Thus, the replicate fractional weights are also computed in two steps. Firstly, the initial replication fractional weight for $w_{ij0}^{*} = 1/m$ is then given by

$$w_{ij0}^{*[k]} = \frac{d_{ij}\left(w_{ij}^{*[k]} / w_{ij}^{*}\right)}{\sum_{l\in A_R} d_{il}\left(w_{il}^{*[k]} / w_{il}^{*}\right)}, \tag{A.2}$$

where w_{ij}^{*} is the fractional weight for FFI defined in (2.6) and $w_{ij}^{*[k]}$ is the k-th replication fractional weight for FFI defined in (A.1). Secondly, the replication fractional weights are adjusted to satisfy the calibration constraints. The calibration equation for replication fractional weights corresponding to (3.5) is then

$$\sum_{i\in A} w_i^{[k]} \left\{ \left(1-\delta_i\right) \sum_{j\in D_i} w_{ij,c}^{*[k]} \mathbf{q}\left(\mathbf{x}_i, y_j\right) \right\} = \sum_{i\in A} w_i^{[k]} \left\{ \left(1-\delta_i\right) \sum_{j\in A_R} w_{ij}^{*[k]} \mathbf{q}\left(\mathbf{x}_i, y_j\right) \right\} \tag{A.3}$$

and $\sum_{j\in D_i} w_{ij,c}^{*[k]} = 1.$ Either regression weighting or entropy weighting can be used to obtain the replication fractional weights satisfying the constraints. Once the replicate fractional weights are obtained, the replicate estimate $\hat{\eta}^{[k]}$ is computed by solving

$$\sum_{i\in A} w_i^{[k]} \left\{ \delta_i U\left(\eta; x_i, y_i\right) + \left(1-\delta_i\right) \sum_{j\in A_R} w_{ijc}^{*[k]} U\left(\eta; x_i, y_j\right) \right\} = 0.$$

The replication variance estimator of $\hat{\eta},$ computed from (3.8), is given by

$$\hat{V}_{rep}\left(\hat{\eta}\right) = \sum_{k=1}^{L} c_k \left(\hat{\eta}^{[k]} - \hat{\eta}\right)^2.$$

Because $\hat{\eta}$ is a smooth function of $\hat{\theta},$ the consistency of $\hat{V}_{rep}(\hat{\eta})$ follows directly from the standard argument of the replication variance estimation (Shao and Tu 1995).

A.2 Proof of Equation (4.5)

Using

$$\frac{g\left(y_j \mid x_i\right)}{g\left(y_j \mid x_k\right)} = \frac{f\left(y_j \mid x_i\right)}{f\left(y_j \mid x_k\right)} \exp\left(\varepsilon \Delta_{ik|j} - \kappa\left(x_i\right) + \kappa\left(x_k\right)\right)$$

where $\Delta_{ik|j} = z\left(x_i, y_j; \theta\right) - z\left(x_k, y_j; \theta\right).$ Based on Taylor linearization and the fact of (4.4), we have

$$\frac{g\left(y_j \mid x_i\right)}{g\left(y_j \mid x_k\right)} \cong \frac{f\left(y_j \mid x_i\right)}{f\left(y_j \mid x_k\right)} \left\{1 + \varepsilon \Delta_{ik|j}\right\}.$$

If we know the true density, the correct fractional weights in (3.3) can be expressed by

$$
\begin{aligned}
w_{ij,g}^{*} &\propto \frac{g\left(y_j \mid x_i\right)}{\sum_{k;\delta_k=1} w_k g\left(y_j \mid x_k\right)} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{g(y_j \mid x_k)}{g(y_j \mid x_i)}\right\}} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \exp\left(\varepsilon \Delta_{ki|j} - \kappa\left(x_i\right) + \kappa\left(x_k\right)\right)\right\}} \\
&\cong \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\left(1+\varepsilon \Delta_{ki|j}\right)\right\}} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\} + \varepsilon \sum_{k;\delta_k=1} w_k \left[\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\left\{z\left(x_k, y_j\right) - z\left(x_i, y_j\right)\right\}\right]} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\} + \varepsilon \lambda^T I_\theta^{-1/2} \sum_{k;\delta_k=1} w_k \left(\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\left\{s\left(x_k, y_j\right) - s\left(x_i, y_j\right)\right\}\right)} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\} + \varepsilon \lambda^T I_\theta^{-1/2} \sum_{k;\delta_k=1} w_k \frac{\partial}{\partial \theta}\left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\}} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\}} \left[1 - \varepsilon \lambda^T I_\theta^{-1/2} \frac{\sum_{k;\delta_k=1} w_k \frac{\partial}{\partial \theta}\left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\}}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\}}\right] \\
&\propto \frac{f\left(y_j \mid x_i\right)}{\sum_{k;\delta_k=1} w_k f\left(y_j \mid x_k\right)} \left[1 - \varepsilon \lambda^T I_\theta^{-1/2} \frac{\frac{\partial}{\partial \theta}\left\{\sum_{k;\delta_k=1} w_k \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\}}{\sum_{k;\delta_k=1} w_k \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}}\right] \\
&= \frac{f\left(y_j \mid x_i\right)}{\sum_{k;\delta_k=1} w_k f\left(y_j \mid x_k\right)} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta}\left\{\frac{1}{\sum_{k;\delta_k=1} w_k \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}}\right\} \\
&= a_{ij} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} a_{ij},
\end{aligned}
$$

where $a_{ij} = f\left(y_j \mid x_i\right) \Big/ \sum_{k;\delta_k=1} w_k f\left(y_j \mid x_k\right)$ and $a_{i+} = \sum_{j;\delta_j=1} a_{ij}$. So, $w^*_{ij,f} = a_{ij} \big/ a_{i+}$ and

$$
\begin{aligned}
w^*_{ij,g} &\cong \frac{a_{ij} + \varepsilon\lambda^T I_\theta^{-1/2} \dfrac{\partial}{\partial\theta} a_{ij}}{a_{i+} + \varepsilon\lambda^T I_\theta^{-1/2} \dfrac{\partial}{\partial\theta} a_{i+}} \\
&= \frac{a_{ij}}{a_{i+}} \left(1 + \frac{\varepsilon\lambda^T I_\theta^{-1/2} \dfrac{\partial}{\partial\theta} a_{ij}}{a_{ij}} \right) \left(\frac{a_{i+}}{a_{i+} + \varepsilon\lambda^T I_\theta^{-1/2} \dfrac{\partial}{\partial\theta} a_{i+}} \right) \\
&\cong \frac{a_{ij}}{a_{i+}} \left(1 + \varepsilon\lambda^T I_\theta^{-1/2} \frac{\partial}{\partial\theta} \log a_{ij} \right) \left(1 - \varepsilon\lambda^T I_\theta^{-1/2} \frac{\partial}{\partial\theta} \log a_{i+} \right) \\
&\cong \frac{a_{ij}}{a_{i+}} + \varepsilon\lambda^T I_\theta^{-1/2} \frac{a_{ij}}{a_{i+}} \left(\frac{\partial}{\partial\theta} \log a_{ij} - \frac{\partial}{\partial\theta} \log a_{i+} \right) \\
&= \frac{a_{ij}}{a_{i+}} + \varepsilon\lambda^T I_\theta^{-1/2} \frac{\partial}{\partial\theta} \left(\frac{a_{ij}}{a_{i+}} \right) \\
&= w^*_{ij,f} + \varepsilon\lambda^T I_\theta^{-1/2} \frac{\partial}{\partial\theta} \left(w^*_{ij,f} \right),
\end{aligned}
$$

which proves (4.5).

A.3 Extension to a non-ignorable missing case

We consider an extension of the proposed method to a non-ignorable missing case. Under the non-ignorable missing assumption, both the conditional model $f\left(y \mid x\right)$ and the response probability model $P\left(\delta = 1 \mid \mathbf{x}, y\right)$ are needed to evaluate the expected estimating function in (4.6). Let the response probability model be given by $Pr\left(\delta_i = 1 \mid \mathbf{x}_i, y_i\right) = \pi\left(\mathbf{x}_i, y_i; \phi\right)$, for some ϕ with a known $\pi(\cdot)$ function. We assume that the parameters are identifiable as discussed in Wang, Shao and Kim (2013).

In PFI, according to Kim and Kim (2012), the MLE $\left(\hat{\theta}, \hat{\phi}\right)$ can be obtained by solving

$$
\sum_{i \in A} w_i \left\{ \delta_i S\left(\theta; \mathbf{x}_i, y_i\right) + \left(1 - \delta_i\right) \sum_{j=1}^{m} w^*_{ij} S\left(\theta; \mathbf{x}_i, y_i^{*(j)}\right) \right\} = 0, \tag{A.4}
$$

and

$$
\sum_{i \in A} w_i \left\{ \delta_i S\left(\phi; \mathbf{x}_i, y_i\right) + \left(1 - \delta_i\right) \sum_{j=1}^{m} w^*_{ij} S\left(\phi; \mathbf{x}_i, y_i^{*(j)}\right) \right\} = 0, \tag{A.5}
$$

where $S\left(\theta; \mathbf{x}, y\right) = \partial \log f\left(y \mid \mathbf{x}; \theta\right) \big/ \partial\theta$, $S\left(\phi; \mathbf{x}, y\right) = \partial \log \pi\left(\mathbf{x}, y; \phi\right) \big/ \partial\phi$, and the fractional weights are given by

$$
w^*_{ij}\left(\theta, \phi\right) = \frac{f\left(y_i^{*(j)} \mid \mathbf{x}_i; \theta\right) \left\{ 1 - \pi\left(\mathbf{x}_i, y_i^{*(j)}, \phi\right) \right\} \Big/ h\left(y_i^{*(j)} \mid \mathbf{x}_i\right)}{\sum_{k=1}^{m} \left[f\left(y_i^{*(k)} \mid \mathbf{x}_i; \theta\right) \left\{ 1 - \pi\left(\mathbf{x}_i, y_i^{*(k)}, \phi\right) \right\} \Big/ h\left(y_i^{*(k)} \mid \mathbf{x}_i\right) \right]}. \tag{A.6}
$$

The solution to (A.4) and (A.5) can be obtained via the EM algorithm. In the EM algorithm, the E-step computes the fractional weights in (A.6) using the current parameter values and the M-step updates the parameter value $\hat{\theta}^{(t+1)}$ and $\hat{\phi}^{(t+1)}$ by solving

$$\sum_{i\in A} w_i \left\{ \delta_i S\left(\theta; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \sum_{j=1}^{m} w_{ij}^{*} \left(\hat{\theta}^{(t)}, \hat{\phi}^{(t)}\right) S\left(\theta; \mathbf{x}_i, y_i^{*(j)}\right) \right\} = 0,$$

and

$$\sum_{i\in A} w_i \left\{ \delta_i S\left(\phi; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \sum_{j=1}^{m} w_{ij}^{*} \left(\hat{\theta}^{(t)}, \hat{\phi}^{(t)}\right) S\left(\phi; \mathbf{x}_i, y_i^{*(j)}\right) \right\} = 0.$$

In the proposed FFI method, the fractional weights are given by

$$\begin{aligned} w_{ij}^{*} &\propto f\left(y_j \mid \mathbf{x}_i, \delta_i = 0; \theta, \phi\right) \Big/ f\left(y_j \mid \delta_j = 1\right) \\ &\propto f\left(y_j \mid \mathbf{x}_i; \theta\right) \left\{1 - \pi\left(\mathbf{x}_i, y_j; \phi\right)\right\} \Big/ f\left(y_j \mid \delta_j = 1\right), \end{aligned}$$

with $\sum_{j;\delta_j=1} w_{ij}^{*} = 1$. Because

$$\begin{aligned} f\left(y_j \mid \delta_j = 1\right) &= \int \pi\left(\mathbf{x}, y_j\right) f\left(y_j \mid \mathbf{x}\right) f(\mathbf{x}) d\mathbf{x} \\ &\cong \sum_{k\in A} w_k \pi\left(\mathbf{x}_k, y_j\right) f\left(y_j \mid \mathbf{x}_k\right). \end{aligned} \tag{A.7}$$

The fractional weights can be computed from

$$w_{ij}^{*} \propto \frac{f\left(y_j \mid \mathbf{x}_i; \theta\right) \left\{1 - \pi\left(\mathbf{x}_i, y_j; \phi\right)\right\}}{\sum_{k\in A} w_k \pi\left(\mathbf{x}_k, y_j; \phi\right) f\left(y_j \mid \mathbf{x}_k; \theta\right)}. \tag{A.8}$$

with $\sum_{j\in A_R} w_{ij}^{*} = 1$.

Thus, we can use the following EM algorithm to obtain the desired parameter estimates.

(I-step) For each missing unit $i \in A_M = \left\{i \in A; \delta_i = 0\right\}$, take m imputed values as $y_i^{(1)}, \ldots, y_i^{(m)}$ from A_R, where $m = r$.

(E-step) The fractional weights are given by

$$w_{ij}^{*(t)} \propto \frac{f\left(y_j \mid \mathbf{x}_i, \hat{\theta}^{(t)}\right) \left\{1 - \pi\left(\mathbf{x}_i, y_j; \hat{\phi}^{(t)}\right)\right\}}{\sum_{k\in A} w_k \pi\left(\mathbf{x}_k, y_j; \hat{\phi}^{(t)}\right) f\left(y_j \mid \mathbf{x}_k; \hat{\theta}^{(t)}\right)}$$

and $\sum_{j=1}^{m} w_{ij}^{*(t)} = 1$.

(M-step) Update the parameter $\hat{\theta}^{(t+1)}$ and $\hat{\phi}^{(t+1)}$ by solving the following imputed score equations,

$$\sum_{i\in A} w_i \left\{ \delta_i S\left(\theta; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \sum_{j\in A_R} w_{ij}^{*(t)} S\left(\theta; \mathbf{x}_i, y_j\right) \right\} = 0,$$

and

$$\sum_{i\in A} w_i \left\{ \delta_i S\left(\phi; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \sum_{j\in A_R} w_{ij}^{*(t)} S\left(\phi; \mathbf{x}_i, y_j\right) \right\} = 0.$$

Note that the I-step does not have to be repeated in the EM algorithm. Once the final parameter estimates are computed, the fractional weights are computed by (A.8), which serve as the selection probabilities for FHDI with a small imputation size m. The same systematic PPS sampling method as discussed in Section 3 can be used to obtain FHDI.

References

Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting process: a large sample study. *The Annals of Statistics*, 10, 1100-1120.

Andridge, R.R. and Little, R.J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78, 40-64.

Beaumont, J.F. and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, 37, 400-416.

Binder, D. and Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.

Chauvet, G., Deville, J.-C. and Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*, 98, 459-471.

Chen, J. and Shao, J. (2001). Jackknife variance estimation for nearest neighbor imputation. *Journal of the American Statistical Association*, 96, 260-269.

Copas, J.B. and Eguchi, S. (2001). Local sensitivity approximations for selectivity bias. *Journal of the Royal Statistical Society, Series B*, 63, 871-895.

Durrant, G.B. (2009). Imputation methods for handling item-nonresponse in practice: methodological issues and recent debates. *International Journal of Social Research Methodology*, 12, 293-304.

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96, 933-944.

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

Haziza, D. (2009). Imputation and inference in the presence of missing data. *Handbook of Statistics*, 29, 215-246.

Kalton, G. and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics, Series A*, 13, 1919-1939.

Kim, J.K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98, 119-132.

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

Kim, J.K., Fuller, W.A. and Bell, W.R. (2011). Variance estimation for nearest neighbor imputation for U.S. census long form data. *Annals of Applied Statistics*, 5, 824-842.

Kim, J.K. and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. Chapman and Hall/CRC.

Kim, J.Y. and Kim, J.K. (2012). Parametric fractional imputation for nonignorable missing data. *Journal of the Korean Statistical Society*, 41, 291-303.

Meng, X.L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 9, 538-573.

Pfeffermann, D. (2011). Modeling of complex survey data: why is it a problem? How should we approach it? *Survey Methodology*, 37, 115-136.

Rao, J.N.K., Yung, W. and Hidiroglou, M.A. (2002). Estimating equations for the analysis of survey data using poststratification information. *The Indian Journal of Statistics, Series A*, 64, 364-378.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-590.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schenker, N. and Welsh, A.H. (1988). Asymptotic results for multiple imputation. *The Annals of Statistics*, 16, 1550-1566.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer.

Sung, Y.J. and Geyer, C.J. (2007). Monte Carlo likelihood inference for missing data models. *The Annals of Statistics*, 35, 990-1011.

Wang, C.-Y. and Pepe, M.S. (2000). Expected estimating equations to accommodate covariate measurement error. *Journal of the Royal Statistical Society, Series B*, 62, 509-24.

Wang, S., Shao J. and Kim, J.K. (2013). An instrument variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*. In press.

Wei, G.C. and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85, 699-704.

White, H. (1982). Maximum likelihood estimation of misspecifed models. *Econometrica*, 50, 1-25.