IOWA STATE UNIVERSITY
Digital Repository

Graduate Theses and Dissertations

Iowa State University Capstones, Theses and Dissertations

2015

Some methods for handling missing data in surveys

Jongho Im
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

 Part of the Statistics and Probability Commons

Recommended Citation
Im, Jongho, "Some methods for handling missing data in surveys" (2015). *Graduate Theses and Dissertations*. 14417.
https://lib.dr.iastate.edu/etd/14417

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

Some methods for handling missing data in surveys

by

Jongho Im

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:

Jae-Kwang Kim, Major Professor

Wayne A. Fuller

Cindy Yu

Yehua Li

Emily Berg

Iowa State University

Ames, Iowa

2015

Copyright © Jongho Im, 2015. All rights reserved.

DEDICATION

I would like to dedicate this thesis to my parents Sukgyu Im and Soonbok Kang. I also dedicate this work to my wife, Minkyung Choi, my daughter, Chaeun Im and my sisters, Eunyoung Im and Eunju Im. They have been my cheerleaders and are very supportive of me throughout my dissertation.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to express a lively sense of gratitude to my main advisor, professor Jae-kwang Kim, who guide me onto the right path and support in regard to research and scholarship. Also, I would like to appreciate committee members, Wayne A. Fuller, Cindy Yu, Yehua Li and Emily Berg, for their encouragement and advice. Particularly, I owe many thanks to professor Wayne A. Fuller for his keen comments in writing of this dissertation.

ABSTRACT

Missing data, or incomplete data, inevitably occurs in many surveys. It is mainly due to nonresponse such that sample units do not fully or partly respond for the survey items. It can be also arisen from sample selection. For example, two-phase sampling can be viewed as a missing data problem in the sense that the study variable is not observed in the first-phase. In truncated data that are intentionally selected by researcher, it will be also missing data problem if we are interested in estimation of non-truncated data properties. Many statistical methods for handling missing data can be categorized into two types based on statistical treatment: one is weighting method and the other is imputation method. The weighting method such as propensity score adjustment that uses response probability as compensation for nonresponse is popular for reducing nonresponse bias. Also, the imputation approach is also prevailed to create complete data for statistical estimation or inference of those imputed data. In this thesis we investigate new statistical methods in both of weighting and imputation methods corresponding to three different missing data situations: (i) propensity score adjustment for nonignorable nonresponse data with several follow-ups, (ii) correlation estimation of singly tuncated bivariate samples and (iii) fractional hot deck imputation for multivariate missing data.

1 OVERVIEW

In this thesis, some statistical methods are newly proposed for handling of missing data. In particular we consider two types of missing data with respect to source of missingness. One is due to nonresponse that is a main reason of missing data in many surveys. The other is a truncated data that are selectively observed during data collection.

Nonresponse can be distinguished between unit nonresponse and item nonresponse based on its usability (Kalton and Kasprzyk, 1986). For unit nonresponse, all survey variables are missing or there is no enough information for statistical estimation or inference. In case of item nonresponse, some survey variables are not usable due to refusal or the result of edit failures. Truncated samples are obtained by excluding certain population values. It can be naturally obtained from restrictions on population or intentionally gathered by purpose of researchers or survey goals.

The proposed methods are particularized corresponding to the response mechanism. Response mechanism can be categorized into three types (Rubin, 1976): (i) MCAR: missing complete at random, (ii) MAR: missing at random and (iii) MNAR: missing not at random. MCAR holds if a unit response is independent both of observed and unobserved survey variables. In MAR, the response is dependent on observed variables and independent of unobserved variables. However, the response also depends on unobserved variables for MNAR. The response mechanism is also called ignorable for MCAR or MAR and nonignorable for MNAR.

In Chapter 2, we develop a propensity score weighting adjustment for nonignorable nonresponse when there are several follow-ups and the final followup sample is also subject to missingness. A method-of-moment type estimator is proposed to estimate parameters of the response probability model. For parameter estimation, we use the generalized method of mo-

ments (GMM) method and variance estimation is also considered with variance estimator of GMM estimator. A limited simulation is conducted to test the proposed method and the results of application for Korean household survey are presented.

Chapter 3 is devoted to estimation of the correlation in a singly truncated bivariate samples. To construct an unbiased estimator of correlation, the joint distribution of bivariate variables are decomposed as a marginal distribution of truncated variable and a conditional distribution obtained from a linear regression model that a truncated variable is used as explanatory variable. Then, an unbiased estimator is obtained by multiplying the regression slope coefficient with variance ratio of two variables. The proposed estimator is compare to estimator obtained from bivariate normal assumption in several simulation results. Also, the proposed method is applied to South Sudan children's anthropometric and nutritional data collected by the World Vision.

In Chapter 4, we propose a fractional hot deck imputation method, under MAR assumption, for handing item nonresponse with arbitrary missing patterns. First, we apply categorization on survey items to construct imputation cells and consider a modified EM algorithm to estimate cell probabilities. After then, the proposed fractional imputation procedure is implemented in spirit of a two-phase stratified sampling in the sense that all possible imputation cells are assigned to the missing items and then imputed values are generated from each imputation cell. Replication variance estimation is discussed and results from two limited simulation studies are presented.

2 PROPENSITY SCORE ADJUSTMENT WITH SEVERAL FOLLOW-UPS

Modified from a paper to be published in *Biometrika*[1]

Jae kwang Kim[2] and Jongho Im[2]

Abstract

Propensity score weighting adjustment is commonly used to handle unit nonresponse. When the response mechanism is nonignorable in the sense that the response probability depends directly on the study variable, a followup sample is commonly used to obtain an unbiased estimator using the framework of two-phase sampling, where the followup sample is assumed to respond completely. In practice, the followup sample is also subject to missingness. We consider propensity score weighting adjustment for nonignorable nonresponse when there are several follow-ups and the final followup sample is also subject to missingness. We propose a method-of-moment estimator for estimating parameters in the response probability. The proposed method can be implemented using the generalized method of moments and a consistent variance estimate can be obtained relatively easily. A limited simulation study shows the robustness of the proposed method. The proposed methods are applied to a Korean household survey of employment.

Key words: Nonignroable nonresponse, Survey sampling, Two-phase sampling, Weighting.

[1]Reprinted with permission of *Biometrika*, **101**, 439–448

[2]Department of Statistics, Iowa State University, Ames, U.S.A.

2.1 Introduction

Propensity score weighting is a popular tool for handling unit nonresponse in survey sampling. Many surveys use propensity score weighting to reduce nonresponse bias (Fuller et al. , 1994; Rizzo et al. , 1996). If the responses are ignorable in the sense of Rubin (1976), then the propensity scores can be estimated consistently and the resulting propensity-score-adjusted estimator is easily constructed. Kott (2006), Kim and Kim (2007), and Kim and Riddles (2012) have investigated some statistical properties of the propensity-score-adjusted estimators under missing at random case. If the responses are not ignorable, however, estimation of the propensity scores is complicated and often requires additional surrogate (Chen et al. , 2008) or instrumental variables (Chang and Kott , 2008; Kott and Chang , 2010) to estimate the model parameters consistently. Generally speaking, parameter estimation in the nonignorable response model can be subject to non-identifiability and often requires additional assumptions (Wang et al. , 2014).

Another way of handling nonignorable response is to use followup samples to obtain further observations. Deming (1953) used two-phase sampling (Neyman , 1938; Hansen and Hurwitz , 1946) to obtain a followup sample in the nonrespondents' stratum and obtained a design-unbiased two-phase sampling estimator where the followup sample is treated as a second-phase sample in the two-phase sampling setup, assuming that the followup sample does not suffer from unit nonresponse. Proctor (1977) used a multinomial distribution to model differential response rate in the followup sample. Drew and Fuller (1980); Drew and Fuller (1981) extended the work of Proctor (1977) and developed a maximum likelihood estimation method for a categorical response variable. Alho (1990) extended the approach of Drew and Fuller to continuous response variables by adopting a logistic regression model for the response probability and proposed a maximum likelihood estimator of the model parameters that maximizes the conditional likelihood for the respondents. Wood et al. (2006) compared Alho (1990)'s conditional likelihood method with a fully parametric unconditional

likelihood of categorized outcome variable which can be estimated by the EM algorithm and a Bayesian approach using the Gibbs sampler.

In practice, we often have nonnegligible nonresponse even after followup attempts. In the Korean Labor Force Survey example discussed in Section 2.6, followup attempts were made up to three times. After the fourth attempt, there are still about 10% nonrespondents in the sample. This paper proposes a calibration weighting method for handling nonresponse after several followups.

There are several advantages of the proposed method. First, it is easy to understand and can incorporate additional auxiliary information. Second, consistent variance estimation is obtained as a by-product of the generalized method of moments estimation. Third, as is demonstrated in Section 2.5.2, it is quite robust against the failure of the assumed response model. Furthermore, it is directly applicable to complex survey sampling.

2.2 Basic Setup

Let $U = \{1, \ldots, N\}$ be the index set of a finite population with known size N and let $A \subset U$ be the original sample obtained from a probability sampling design. Let y_i be the study variable. Let d_i be the sampling weight assigned to unit i in the sample so that the resulting estimator

$$\hat{Y}_d = \sum_{i \in A} d_i y_i$$

is unbiased for the total $Y = \sum_{i=1}^{N} y_i$.

Now, suppose that the original sample is not fully observed and there are followups to increase the number of respondents. Let $A_1 \subset A$ be the set of initial respondents who provided answers at the initial contact. Suppose that there are $T-1$ followups made to those who remain nonrespondents in the survey. Let $A_2 \subset A$ be the set of respondents who provided answers through the first followup. By definition, A_2 contains those who provided answers in the initial contact. Thus, $A_1 \subset A_2$. Similarly, we can define A_3 be the set of respondents who provided

answers through the second followup. Continuing the process, we can define $A_1, \ldots, A_T$ such that

$$A_1 \subset \cdots \subset A_T.$$

Followup can be also called call-back. Suppose that there are T attempts, or $T-1$ followups, to obtain the survey response y_i and let δ_{it} be the response indicator function for y_i at the t^{th} attempt. If an unit never responds to all attempts, the unit is called hardcore nonrespondent (Drew and Fuller , 1980). Using the definition of A_t, we can write $\delta_{it} = 1$ if $i \in A_t$ and $\delta_{it} = 0$ otherwise.

When the study variable y is categorical with K categories, Drew and Fuller (1980) proposed using a multinomial distribution with $T \times K + 1$ cells, with cell probabilities defined by

$$\pi_0 = (1-\gamma) + \gamma \sum_{k=1}^{K} (1-p_k)^T f_k, \pi_{tk} \;\; = \;\; \gamma(1-p_k)^{t-1} p_k f_k, \; t = 1, \ldots, T, \; k = 1, \ldots, K,$$

where p_k is the response probability for category k, f_k is the population proportion such that $\sum_{k=1}^{K} f_k = 1$ and $1-\gamma$ is a proportion of hardcore nonrespondents. Thus, π_{tk} means the response probability that an individual in category k will respond at the t^{th} contact and π_0 is the probability that an individual will not have responded after T trials. That is, π_0 includes both hardcore nonrespondents and others who do not respond during T attempts. Under simple random sampling, the maximum likelihood estimator of the parameter can be obtained by maximizing the log-likelihood

$$\log L \;\; = \;\; n_0 \log \pi_0 + \sum_{t=1}^{T} \sum_{k=1}^{K} n_{tk} \log \pi_{tk},$$

where n_{tk} is the number of elements in the k^{th} category responding on the t^{th} contact and n_0 is the number of individual who did not respond up to the T^{th} contact. Drew and Fuller (1981) further extended the results to complex survey sampling.

Alho (1990) considered the same problem with a continuous y variable under the simple random sampling. Alho (1990) defined p_{it} to be the conditional probability of $\delta_{it} = 1$,

conditional on y_i and $\delta_{i,t-1}=0$, and used the logistic regression model

$$p_{it}=\text{pr}(\delta_{it}=1\mid\delta_{i,t-1}=0,x_i,y_i)=\frac{\exp\left(\alpha_t+x_i\phi_1+y_i\phi_2\right)}{1+\exp\left(\alpha_t+x_i\phi_1+y_i\phi_2\right)},\quad t=1,\ldots,T, \tag{2.1}$$

for the conditional response probability with $\delta_{i0}\equiv 0$, where x_i is the auxiliary variable that is available throughout the sample.

To estimate the parameters in (2.1), Alho (1990) also assumed that $(\delta_{i1},\delta_{i2}-\delta_{i1},\ldots,\delta_{iT}-\delta_{i,T-1},1-\delta_{iT})$ follows from a multinomial distribution with parameter $(\pi_{i1},\pi_{i2},\ldots,\pi_{iT},1-\sum_{t=1}^{T}\pi_{it})$ where $\pi_{it}=\text{pr}\left(\delta_{i,t-1}=0,\delta_{it}=1\mid x_i,y_i\right)$. Thus, we can write $\pi_{it}=p_{it}\prod_{k=1}^{t-1}(1-p_{ik})$. Under this setup, Alho (1990) considered maximizing the conditional likelihood,

$$\begin{aligned}L_c(\phi) &= \prod_{\delta_{iT}=1}\left\{\text{pr}\left(\delta_{i1}=1\mid x_i,y_i,\delta_{iT}=1\right)^{\delta_{i1}}\prod_{t=2}^{T}\text{pr}\left(\delta_{i,t-1}=0,\delta_{it}=1\mid x_i,y_i,\delta_{iT}=1\right)^{\delta_{it}}\right\}\\ &= \prod_{\delta_{iT}=1}\left(\frac{\pi_{i1}}{1-\pi_{i,T+1}}\right)^{\delta_{i1}}\prod_{t=2}^{T}\left(\frac{\pi_{it}}{1-\pi_{i,T+1}}\right)^{\delta_{it}-\delta_{i,t-1}}\end{aligned} \tag{2.2}$$

where $\pi_{i,T+1}=1-\sum_{t=1}^{T}\pi_{it}$. For identifiability, Alho (1990) imposed

$$\sum_{i\in A-A_{t-1}}\delta_{it}\exp\left(-\alpha_t-\phi_1x_i-\phi_2y_i\right)=n-(n_1+\cdots+n_t),\quad t=1,\ldots,T. \tag{2.3}$$

The equation (2.3) defines α_t given ϕ.

Alho's method used $\sum_{t=1}^{T}\hat{\pi}_{it}=1-\hat{\pi}_{i,T+1}$ to compute the propensity-score-adjusted estimator

$$\hat{\theta}_{\text{PSA}}=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_{iT}}{(1-\hat{\pi}_{i,T+1})}y_i. \tag{2.4}$$

Alho did not discuss variance estimation for (2.4). Furthermore, Alho's method does not make use of the auxiliary variable x_i in the nonrespondents.

2.3 Conditional maximum likelihood method

We first consider a generalization of Alho (1990)'s method for parameter estimation in the propensity score model. The basic idea is to maximize the conditional likelihood using the set of respondents, those with $\delta_{iT}=1$, for whom the response probability is reversed in the

sense that, instead of the original probability in (2.1), the conditional probability of $\delta_{i,t-1} = 1$ given that $\delta_{it} = 1$ is considered. The conditional likelihood was also considered by Tang et al. (2003) and Pfeffermann and Sikov (2011) for the special case $T = 1$, i.e., no followup.

The approach based on conditional likelihood consists of two steps. In the first step, the reverse conditional probability $q_{it} = \text{pr}\left(\delta_{it} = 1 \mid \delta_{i,t+1} = 1, x_i, y_i\right)$ is derived using Bayes' formula from the assumed response model. That is, we can obtain $q_{it} = O_{it}/(1 + O_{it})$ where

$$\begin{aligned} O_{it} &\equiv \frac{\text{pr}(\delta_{it} = 1 \mid x_i, y_i, \delta_{i,t+1} = 1)}{\text{pr}(\delta_{it} = 0 \mid x_i, y_i, \delta_{i,t+1} = 1)} \\ &= \frac{\text{pr}(\delta_{it} = 1, \delta_{i,t+1} = 1 \mid x_i, y_i)}{\text{pr}(\delta_{it} = 0, \delta_{i,t+1} = 1 \mid x_i, y_i)} \\ &= \frac{\text{pr}(\delta_{i,t+1} = 1 \mid x_i, y_i, \delta_{i,t} = 1)}{\text{pr}(\delta_{i,t+1} = 1 \mid x_i, y_i, \delta_{i,t} = 0)} \frac{\text{pr}(\delta_{it} = 1 \mid x_i, y_i)}{\text{pr}(\delta_{it} = 0 \mid x_i, y_i)} \\ &= \frac{1}{p_{i,t+1}} \frac{\tilde{\pi}_{it}}{1 - \tilde{\pi}_{it}} \end{aligned}$$

and $\tilde{\pi}_{it} = \sum_{j=1}^{t} \{p_{ij} \prod_{k=1}^{j-1} (1 - p_{ik})\} = \sum_{j=1}^{t} \pi_{ij}$. Thus, we can express q_{it} as a function of $\alpha = (\alpha_1, \ldots, \alpha_t)$ and ϕ in (2.1).

In the second step, the parameter estimator is obtained by maximizing the conditional likelihood based on the reverse conditional probability at time t,

$$L_t(\alpha, \phi) = \prod_{i \in A_{t+1}} q_{it}^{\delta_{it}} \left(1 - q_{it}\right)^{1 - \delta_{it}},$$

where q_{it} is a function of $\alpha = (\alpha_1, \ldots, \alpha_t)$ and ϕ in (2.1). For samples obtained from an unequal probability sampling design, we can consider maximizing the pseudo conditional log-likelihood function

$$l_c(\alpha, \phi) = \sum_{t=1}^{T-1} \sum_{i \in A} d_i \delta_{i,t+1} \left\{\delta_{it} \log(q_{it}) + (1 - \delta_{it}) \log(1 - q_{it})\right\}. \tag{2.5}$$

Under simple random sampling, the conditional log-likelihood in (2.5) is essentially the same as that of Alho (1990), which is presented in (2.2). Given the identifiability constraint (2.3), we may add another constraint to incorporate the observed auxiliary information outside A_t,

$$\sum_{i \in A} d_i \frac{\delta_{iT}}{1 - \pi_{i,T+1}} x_i = \sum_{i \in A} d_i x_i. \tag{2.6}$$

Incorporating the constraints into the propensity score estimation is equivalent to finding the solution that is the stationary point of the Lagrangian function

$$L(\alpha, \phi, \lambda) \quad = \quad l(\alpha, \phi) + \lambda^{\mathrm{T}} g(\alpha, \phi),$$

where $g(\alpha, \phi)$ are the constraint functions in (2.3) and (2.6). Once the parameters are estimated, our final propensity-score-adjusted estimator is

$$\hat{Y}_{\mathrm{PSA}} = \sum_{i \in A} d_i \frac{\delta_{iT}}{(1 - \hat{\pi}_{i,T+1})} y_i. \tag{2.7}$$

Although the conditional maximum likelihood method can lead to efficient propensity-score-adjusted estimator, the computation for constrained optimization is complicated and consistent variance estimation would require very tedious Taylor linearization. Furthermore, maximum likelihood estimation is often sensitive to departures from the assumed response model, as demonstrated in Section 2.5.

2.4 Calibration method

In this section, we propose an approach based on moment conditions to estimate the model parameters in the conditional response model. For simplicity of presentation, we first let $T = 2$, where there is only one followup from the set of nonrespondents in the original sample.

From the set of initial respondents with $\delta_{i1} = 1$, we have

$$E\left\{\sum_{i \in A} d_i \frac{\delta_{i1}}{p_{i1}}(1, x_i, y_i)\right\} = (N, X, Y), \tag{2.8}$$

where $p_{i1} = \mathrm{pr}(\delta_{i1} = 1 \mid x_i, y_i)$ and (X, Y) are population totals, $(X, Y) = \sum_{i=1}^{N}(x_i, y_i)$. Also, from the set of respondents at time $t = 2$, we have

$$E\left\{\sum_{i \in A} d_i \delta_{i1}(1, x_i, y_i) + \sum_{i \in A} d_i \frac{(1 - \delta_{i1})\delta_{i2}}{p_{i2}}(1, x_i, y_i)\right\} = (N, X, Y), \tag{2.9}$$

where $p_{i2} = \mathrm{pr}(\delta_{i2} = 1 \mid x_i, y_i, \delta_{i1} = 0)$. Thus, combining (2.8) and (2.9), we obtain

$$\sum_{i \in A} d_i \frac{\delta_{i1}}{p_{i1}}(1, x_i, y_i) = \sum_{i \in A} d_i \delta_{i1}(1, x_i, y_i) + \sum_{i \in A} d_i \frac{(1 - \delta_{i1})\delta_{i2}}{p_{i2}}(1, x_i, y_i), \tag{2.10}$$

which can be used to as calibration equations to estimate the model parameters in the response model. Under Alho's model (2.1), equation (2.10) reduces to

$$\sum_{i\in A} d_i\delta_{i1}\{1+\exp(-\alpha_1-\phi_1 x_i-\phi_2 y_i)\}(1,x_i,y_i)$$
$$=\sum_{i\in A} d_i\delta_{i1}(1,x_i,y_i)+\sum_{i\in A} d_i(1-\delta_{i1})\delta_{i2}\{1+\exp(-\alpha_2-\phi_1 x_i-\phi_2 y_i)\}(1,x_i,y_i), \qquad (2.11)$$

which is a system of $p+q+1$ equations with $p+q+2$ parameters, where $p=\dim(x)$ and $q=\dim(y)$. To uniquely determine the parameters, we assume N to be known or at least estimated by $\hat{N}=\sum_{i\in A} d_i$, which is equivalent to adding another condition

$$\sum_{i\in A} d_i=\sum_{i\in A} d_i\delta_{i1}\left\{1+\exp(-\alpha_1-\phi_1 x_i-\phi_2 y_i)\right\}. \qquad (2.12)$$

Thus, solving (2.11) and (2.12) simultaneously, we can obtain the parameters of the response mechanism consistently. The final estimator of Y is then

$$\hat{Y}=\sum_{i\in A} d_i\delta_{i1}\left\{1+\exp(-\hat{\alpha}_1-\hat{\phi}_1 x_i-\hat{\phi}_2 y_i)\right\}y_i$$

which also equals

$$\sum_{i\in A} d_i\delta_{i1}y_i+\sum_{i\in A} d_i(1-\delta_{i1})\delta_{i2}\left\{1+\exp(-\hat{\alpha}_2-\hat{\phi}_1 x_i-\hat{\phi}_2 y_i)\right\}y_i$$

by construction. The proposed method can be called a method-of-moments approach or a calibration equation approach because the weights are constructed to satisfy some moment conditions. When the response mechanism is ignorable, i.e., $\phi_2=0$, the calibration equation approach is popular in the propensity score weighting literature (Folsom , 1991; Ianncchione et al. , 1991; Fuller et al. , 1994; Kott , 2006; Kim and Riddles , 2012).

We now discuss an extension of the calibration equation method to the case with $T\geq 2$. The calibration equations can be written as

$$\sum_{i\in A} d_i\delta_{i,t-1}(1,x_i,y_i)+\sum_{i\in A} d_i(1-\delta_{i,t-1})\frac{\delta_{it}}{p_{it}}(1,x_i,y_i)=(N,X,Y), \qquad (2.13)$$

for $t=1,\ldots,T$, with $\delta_{i0}=0$, and

$$\sum_{i\in A} d_i(1,x_i)=(N,X) \qquad (2.14)$$

where $(\alpha_1, \ldots, \alpha_T, \phi_1, \phi_2)$ are the parameters in the response model and (N, X, Y) are unknown parameter to be determined. The left-hand side of (2.13) is an unbiased estimator of (N, X, Y) obtained from the sample at time t, respectively. Combining (2.13) and (2.14), we have $T + 2p + 2q + 1$ parameters with $(p + q + 1) \times T + p + 1$ equations. If $T = 2$ and $p = 0$, then the number of equations equals the number of parameters. Otherwise, we have more estimating equations than the parameters. If (N, X) is known, then we can still use the same equations but the number of parameters reduces to $T + p + 2q$.

When we have more equations than parameters, we can apply the generalized method of moments technique to compute the estimates. Writing $\eta = (\alpha_1, \cdots, \alpha_T, \phi, X, Y)$, the generalized method of moments estimates $\hat{\eta}$ can be obtained by minimizing

$$Q = \hat{U}(\eta)^{\mathrm{T}} \hat{V}\{\hat{U}(\eta)\}^{-1} \hat{U}(\eta) \tag{2.15}$$

where $\hat{U}(\eta)$ is the system of estimating equations derived from (2.13) and (2.14) and $\hat{V}\{\hat{U}(\eta)\}$ is a design-consistent variance estimator of $\hat{U}(\eta)$ for a fixed value of η. Computational details are presented in Appendix. Under some regularity conditions, the generalized method of moments estimator of η is approximately unbiased with the asymptotic variance estimated by

$$\hat{V}(\hat{\eta}) = \left[\hat{\tau} \hat{V}\{\hat{U}(\hat{\eta})\}^{-1} \hat{\tau}^{\mathrm{T}}\right]^{-1}, \tag{2.16}$$

where $\hat{\tau} = \partial \hat{U} / \partial \eta^{\mathrm{T}}$ evaluated at $\eta = \hat{\eta}$.

2.5 Simulation study

2.5.1 Simulation One

In this section, we presents the results from two simulation studies. The first compares the statistical efficiency of the estimators and the second compares the robustness of the estimators under nonignorable nonresponse with missing data in the followup samples.

In the first simulation study, we generated $B = 2,000$ Monte Carlo samples of size $n = 600$ with variables (x_i, y_i) from $x_i \sim$ Uniform$(-2, 2)$, Model 1: $y_i = 0.8 + x_i + e_i$ and Model 2:

Table 2.1. Performance of three estimators when $n = 600,\ T = 2$. The biases, standard errors, and root mean squared errors are multiplied by 100.

		Correctly specified model			Mis-specified model		
Model	Estimator	Bias	SE	RMSE	Bias	SE	RMSE
	Alho	0.20	6.48	6.48	−14.92	6.08	16.11
Model 1	CK	−0.51	6.00	6.02	−2.97	5.82	6.54
	CAL	0.77	5.92	5.97	−0.59	5.74	5.77
	Alho	0.15	5.00	5.00	−14.21	5.29	15.16
Model 2	CK	1.74	7.75	7.94	−7.75	8.54	11.54
	CAL	0.23	5.57	5.57	−1.19	4.90	5.04

CAL, the proposed calibration estimator; CK, Chang and Kott's estimator; RMSE, root mean squared error; SE, standard error.

$y_i = 0.4x_i + 0.6x_i^2 + e_i$, where $e_i \sim N(0, 1/2)$. The population correlation between x and y are about 0.85 and 0.42 for Model 1 and Model 2, respectively. We assumed one callback to obtain the followup samples and also assumed the following conditional response model

$$p_{it} \equiv \text{pr}(\delta_{it} = 1 \mid x_i, y_i, \delta_{i,t-1} = 0) = \{1 + \exp(-\alpha_t - \phi y_i)\}^{-1}, t = 1, 2, \tag{2.17}$$

where $(\alpha_1, \alpha_2, \phi) = (-1, 0.5, 1)$. Note that x_i is the nonresponse instrumental variable in this setup because it is conditionally independent of $(\delta_{i1}, \delta_{i2})$ given y_i.

We computed three estimators of $\theta = E(Y)$: Alho's estimator obtained from (2.7), Chang and Kott's estimator (Chang and Kott , 2008) and the proposed calibration equation estimator obtained by minimizing (2.15). Drew and Fuller (1980) method was not considered in the simulation because it is applicable only to categorial responses. Chang and Kott's estimator is

$$\hat{\theta}_{\text{CK}} = \frac{\sum_{i=1}^n \delta_{iT} y_i / \hat{\pi}_i}{\sum_{i=1}^n \delta_{iT} / \hat{\pi}_i}$$

where $\hat{\pi}_i$ is obtained by solving $\sum_{i=1}^n (\delta_{iT}/\pi_i - 1)(1, x_i) = 0$ with $\pi_i = \{1 + \exp(-\alpha^* - \phi^* y_i)\}^{-1}$.

Table 2.1 presents the biases, standard errors, and the root mean squared errors of the three estimators under two models with correctly specified response model. All estimators are nearly unbiased except for Chang and Kott's estimator when the linear relationship between y and x

does not hold. Under model 1, both Chang and Kott's and the proposed calibration equation estimator are more efficient than Alho's estimator because they directly use the calibration equation, which leads to efficient estimation under a linear relationship. Under Model 2, the linear relationship does not hold and Alho's estimator is slightly more efficient than the proposed estimator because it is based on the maximum likelihood approach. However variance estimation is not easy for Alho's estimator and, as can be seen in Section 2.5.2, it is not robust against failure of the assumed response model. Chang and Kott's estimator is very unstable results when the linear relationship between y and x does not hold. Variance estimation of the calibration equation estimator is computed by (2.16). The relative biases of the variance estimators of calibration equation estimator are less than 5% in both models and are not presented here.

2.5.2 Simulation Two

Under the setup of Simulation One, we considered another type of conditional response model to check the robustness of estimators against mis-specification of this model. In this simulation, we used the distribution as the true response model

$$p_{it} \equiv \text{pr}(\delta_{it} = 1 \mid x_i, y_i, \delta_{i,t-1} = 0) = \frac{\Gamma(\alpha_t + \phi)}{\Gamma(\alpha_t)\Gamma(\phi)} z_i^{\alpha_t - 1}(1 - z_i)^{\phi - 1},$$

where $(\alpha_1, \alpha_2, \phi) = (1, 0.5, 3)$ and $z_i = y_i^2/(1 + y_i^2)$. We still used the logistic regression model (2.17) as the working model for the response mechanism.

Table 2.1 presents the biases, standard errors, and the root mean squared errors of the point estimators under the mis-specified model. Alho's estimator shows significant biases under both models under the incorrect response model because maximum likelihood estimation is sensitive to departures from the assumed model. Because the proposed calibration equation estimator uses only calibration estimation to estimate the model parameters, the estimated response probability under the incorrect response model still satisfies (2.10) by construction. Thus, as can be seen in Table 2.1, the calibration equation estimator shows the smallest root

Table 2.2 Realized responses in 2009 Korean Local-Area Labor Force survey

Status	$T=1$	$T=2$	$T=3$	$T=4$	No response
Employment	81,685	46,926	28,124	15,992	
Unemployment	1,509	948	597	352	32,350
Not in labor force	57,882	32,308	19,086	10,790	

mean squared errors in both models even when the response mechanism is mis-specified.

2.6 Application

In this section, we present an application of the proposed method to the 2009 Korean Local-Area Labor Force Survey. This large-scale labor force survey is designed to get improved local-area level estimates. Its samples are a stratified two-stage cluster sample of households in Korea, the primary sampling unit is the segment and the secondary sampling unit is the household. The segments consist of about 30 – 80 households, and are selected with probability-proportional-to-size sampling within each stratum, where the measure of size for the sampling is the number of households based on recent Census information. In the 2009 Korean Local-Area Labor Force Survey data, $n = 157,205$ sample households were contacted with up to four followups. Table 2.2 displays the realized number of respondents for each of the followup attempts.

We assume the conditional response model

$$\text{pr}(\delta_{it} = 1 \mid \delta_{i,t-1} = 0, y_i) = \{1 + \exp(-\alpha_t - \phi y_i)\}^{-1}, \tag{2.18}$$

for some (α_t, ϕ), where y_i is the number of unemployed family members in the ith household. We are interested in estimating θ_1, θ_2 and θ_3, which denote the proportion of employment, unemployment and not in labor force, respectively. Note that $\theta_3 = 1 - \theta_1 - \theta_2$ and so we report results for θ_1 and θ_2 only. Under the assumed response model (2.18), we obtained four different estimates. The first is the naive estimator that is computed as the simple mean of the

Table 2.3. Estimates for labor force in 2009 Korean Local-Area Labor Force survey

	θ_1		θ_2	
Method	Estimates($\times 10^2$)	Standard error($\times 10^4$)	Estimates($\times 10^2$)	Standard error($\times 10^4$)
Naive	58.31	11.05	1.15	2.00
Alho	58.30	10.94	1.19	2.56
Drew & Fuller	58.47	10.90	1.19	2.46
Calibration	58.35	11.05	1.19	2.32

respondents without making any adjustment. The other estimates are computed using Alho (1990)'s method, Drew and Fuller (1980)'s method and our proposed method. Alho's method uses the model in (2.18). In computing Alho's method, we use the conditional probability model (2.18).

Table 2.3 presents the estimates. In Table 2.3, the three more sophisticated methods produce slightly larger estimates for the unemployment rate than the naive method, which implies that the missing rate is higher for unemployed people. Since we do not have any an auxiliary variable that is available throughout the sample in this survey, the three methods show similar results.

Acknowledgement

We thank Dr. Seo-young Kim, two referees, the associate Editor, and editor for constructive comments. The research of the first author was supported by a Cooperative Agreement between the U.S. Department of Agriculture Natural Resources Conservation Service and Iowa State University. The research of the second author was partially supported by Korea National Research Foundation for a study of statistical inference for nonlinear long-memory time series data.

Appendix

Computation for generalized method of moments

To discuss the generalized method of moments computation in Section 2.4, suppose for simplicity that $p = 1$ and $q = 1$ such that the conditional response probability will be $p_{it} = g(x_i, y_i; \alpha_t, \phi_1, \phi_2)$ for $t = 1, \ldots, T$. Let $X = \sum_{i=1}^N x_i$ and $Y = \sum_{i=1}^N y_i$. Writing $\hat{\theta}_t(x) = \sum_{i \in A} d_i \{\delta_{i,t-1} + (1 - \delta_{i,t-1})\delta_{it}/p_{it}\} x_i$, the calibration equation can be expressed as

$$\left(\hat{\theta}_t(1), \hat{\theta}_t(x), \hat{\theta}_t(y)\right) = (N, X, Y) \ \ (t = 1, \ldots, T), \ \ \left(\hat{\theta}_{\text{HT}}(1), \hat{\theta}_{\text{HT}}(x)\right) = (N, X).$$

Thus, writing $\theta = (\alpha_1, \cdots, \alpha_T, \phi_1, \phi_2)$, we have estimating formation

$$U(\theta, Y) = \sum_{i \in A} d_i \left\{u_i(1), u_i(x_i), u_i(y_i)\right\}^{\text{T}},$$

where $u_i(z_i) = \{u_{i1}(z_i), \ldots, u_{iT}(z_i)\}$ with $u_{it}(z_i) = \{\delta_{i,t-1} + (1 - \delta_{i,t-1})\delta_{it}/p_{it}(\theta) - 1\} z_i$ for $z_i = 1$ or $z_i = x_i$ and $u_{it}(z_i) = \{\delta_{i,t-1} + (1 - \delta_{i,t-1})\delta_{it}/p_{it}(\theta)\} z_i - n^{-1}Y/d_i$ for $z_i = y_i$ $(t = 1, \ldots, T)$.

Writing $\eta = (\theta, Y)$, the covariance matrix of $U(\eta)$ is easily computed for a fixed value of η by ignoring the randomness of δs. The optimal value of η can be obtained by minimizing Q with respect to η

$$Q = U(\eta)^{\text{T}} V \left(U\right)^{-1} U(\eta). \tag{2.19}$$

We now discuss how to compute the covariance matrix $V(U)$ in (2.19). Let an design-unbiased estimator of $\hat{\theta}_{\text{HT}}(y) = \sum_{i \in A} d_i y_i$ be of the form $\hat{V} = \sum_{i \in A} \sum_{j \in A} \Delta_{ij} y_i y_j$. To estimate the variance estimator of $\hat{\theta}_t(x) = \sum_{i \in A} d_i \eta_{it}$, where $\eta_{it} = \delta_{it} + (1 - \delta_{i,t-1})\delta_{it} x_i / p_{it}$, the naive variance estimator

$$\hat{V}_{\text{naive},t} = \sum_{i \in A} \sum_{j \in A} \Delta_{ij} \eta_{it} \eta_{jt}$$

can be constructed, where $\Delta_{ij} = (\pi_{ij} - \pi_i\pi_j)/(\pi_{ij}\pi_i\pi_j)$. To see its unbiasedness, note that

$$\begin{aligned}
E\left(\hat{V}_{\text{naive},t}\right) &= E\left\{\sum_{i\in A}\sum_{j\in A}\Delta_{ij}E\left(\eta_{it}\eta_{jt} \mid A_{t-1}\right)\right\} \\
&= E\left\{\sum_{i\in A}\sum_{j\in A}\Delta_{ij}E\left(\eta_{it} \mid A_{t-1}\right)E\left(\eta_{jt} \mid A_{t-1}\right)\right\} + E\left\{\sum_{i\in A}\sum_{j\in A}\Delta_{ij}\text{cov}\left(\eta_{it},\eta_{jt} \mid A_{t-1}\right)\right\} \\
&= E\left(\sum_{i\in A}\sum_{j\in A}\Delta_{ij}x_ix_j\right) + E\left\{\sum_{i\in A}\Delta_{ii}\text{var}\left(\eta_{it} \mid A_{t-1}\right)\right\} \\
&= \text{var}\left(\sum_{i\in A}d_ix_i\right) + E\left\{\sum_{i\in A}\left(\frac{1-\pi_i}{\pi_i^2}\right)(1-\delta_{i,t-1})(p_{it}^{-1}-1)x_i^2\right\}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\text{var}\left\{\hat{\theta}_t(x)\right\} &= \text{var}\left[E\left\{\hat{\theta}_t(x) \mid A_{t-1}\right\}\right] + E\left[\text{var}\left\{\hat{\theta}_t(x) \mid A_{t-1}\right\}\right] \\
&= \text{var}\left(\sum_{i\in A}d_ix_i\right) + E\left\{\sum_{i\in A}\pi_i^{-2}\left(1-\delta_{i,t-1}\right)\left(p_{it}^{-1}-1\right)x_i^2\right\}.
\end{aligned}$$

Thus, ignoring the finite population correction term, the naive variance estimator is unbiased.

3 CORRELATION ESTIMATION WITH SINGLY TRUNCATED VARIABLES

Submitted in *Statistical Methods in Medical Research*

Jongho Im[1], Eunyong Ahn[2], Namseon Beck[3], Jae-Kwang Kim[1] and Taesung Park[4]

Abstract

Correlation coefficient estimates are often attenuated for truncated samples. Motivated from a real data in South Sudan, we consider correlation coefficient estimation in a singly truncated bivariate data. By considering a linear regression model where a truncated variable is used as an explanatory variable, a consistent estimate for the slope can be obtained from the ordinary regression method. A consistent estimator of correlation coefficient is then obtained by multiplying the regression slope coefficient with the variance ratio of the two variables. Two simulation studies were conducted to check the performance of the proposed correlation estimator. The simulation study shows that the proposed estimator is nearly unbiased even under non-normal error distributions. The proposed method is applied to South Sudan children's anthropometric and nutritional data collected by the World Vision.

Key words: A Pearson's correlation coefficient, outcome dependent sampling, selection bias, truncated regression.

[1]Department of Statistics, Iowa State University, Ames, U.S.A.
[2]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea
[3]International Department, Medair, Echublens, Switzerland
[4]Department of Statistics, Seoul National University, Seoul, South Korea

3.1 Introduction

A singly truncated bivariate data are often collected in public health data. For example, we consider in this paper mid-upper arm circumference (MUAC) and weight-for-height Z-score (WHZ) which are often used to determine malnutrition status of infants or young children. In many cases, children are firstly selected based on MUAC criterion and then WHZ is also measured in the following examinations. On the basis of this process, we have bivariate data (MUAC, WHZ) when MUAC is right truncated and WHZ is not truncated but associated with MUAC. Considering that MUAC is more easily measured than WHZ in the field study at Africa, we are interested in finding how much truncated MUAC is correlated with WHZ. However, the naive approach of computing the sample correlation coefficient estimate is generally biased in a singly truncated data.

One possible way to get an unbiased estimate is to use the maximum likelihood estimator or moment-based estimator under the bivariate normality assumption. In case of singly truncated bivariate data, Aitkin (1964) and Arnold, Beaver and Meeker (1993) proposed a correlation estimator by assuming a bivariate normal distribution. Early works of Cohen (1955), Rosenbaum (1961) and Rao, Garg and Li (1968) were also limited to a bivariate normal distribution. Arismendi (2013) intensively studied the calculation of truncated moments with a multivariate normal distribution.

Here we use a linear regression model rather than the bivariate normal distribution for correlation estimation. To consider more general bivariate structures, we decompose the joint probability of two variables as the product of the conditional distribution and the marginal distribution. The conditional distribution is assumed to be a normal linear regression model and the marginal distribution is not necessary to be normal. The linear regression model is constructed so that the singly truncated variable is used as the explanatory variable in the regression model. Then, the regression coefficients are consistently estimated even if the explanatory variable is truncated. This desirable result provides an unbiased Pearson's correlation

coefficient estimator by incorporating variance ratio of the two variables.

In Section 3.2, the basic setup is introduced with model assumptions. In Section 3.3, a correlation estimator is proposed in terms of the product of slope coefficient and variance ratio of variables. Results from a simulation study are presented in Section 3.4 and an application of proposed method to South Sudan children's anthropometric and nutrition data is conducted in Section 3.5. Concluding remarks are made in Section 3.6.

3.2 Basic setup

Suppose that we have a sample of two random variables (x, y), where x is right truncated, $x_i \leq x_c$, with known x_c, in the sampling procedure. The truncated sample can be understood as a sample from two-phase sampling, where the first-phase sample is a random sample and the second-phase sample is a outcome-dependent sample in the sense that the cases with $x_i > x_c$ is not selected in the second-phase sampling. Let δ_i be the sampling indicator for the second-phase sampling where $\delta_i = 1$ represents the case when the sample is included in the final sample and $\delta_i = 0$ otherwise. Outcome-dependent sampling in the context of two-phase sampling is very popular in epidemiology. For example, see Breslow and Cain (1988), Kalbfleisch and Lawless (1988), Wild (1991) and Breslow and Holubkov (1997). Unlike the classical two-phase sampling, we do not have the first-phase sample and only have the second-phase sample available for analysis.

If our goal is to estimate the regression parameters in the regression of y on x, we have only to use the final sample for regression analysis without worrying about the sampling procedure. To see this, note that

$$f(y \mid x, \delta = 1) = f(y \mid x) \frac{P(\delta = 1 \mid x, y)}{\int P(\delta = 1 \mid x, y) f(y \mid x) dY}. \tag{3.1}$$

If the sampling mechanism is a function of x, then $P(\delta = 1 \mid x, y) = P(\delta = 1 \mid x)$ and (3.1) reduces to

$$f(y \mid x, \delta = 1) = f(y \mid x). \tag{3.2}$$

Thus, we can safely ignore the sampling mechanism and apply the standard regression techniques to the final data. That is, in estimating the regression parameters, the sampling mechanism is non-informative in the sense of Skinner (1994) and Pfeffermann and Sverchkov (1999). However, for estimating the correlation coefficient, the sampling mechanism becomes informative because $f(x, y \mid \delta = 1) \neq f(x, y)$. Thus, we cannot directly use the standard method to estimate the correlation coefficient.

Early works of correlation estimation with singly truncated data were mostly based on bivariate normal assumption. Aitkin (1964) provided correlation estimator in terms of Mill's ratio,

$$\hat{\rho} = \frac{rR(x_c)}{\sqrt{R(x_c)^2 + (1 - r^2)(x_c R(x_c) - 1)}},$$

where $R(x_c) = \exp(x_c^2/2) \int_{-\infty}^{x_c} \exp(-t^2/2)dt$ is Mill's ratio and r is a sample correlation coefficient of truncated samples. His work is limited to the standard bivariate normal distribution.

Arnold, Beaver and Meeker (1993) extended Aitkin (1964)'s work to a general bivariate normal distribution with mean vector (μ_x, μ_y), variance vector (σ_x^2, σ_y^2) and correlation ρ. From the truncated variable x, they derived the distribution of y in terms of skew normal distribution. They expressed three moments of untruncated variable y of $E(y)$, var(y) and Skewness(y) with three unknown parameters, μ_y, σ_y^2 and ρ, and then obtained the method of moments estimates for the unknown parameters by solving three moment equations. Also they provided likelihood function of y including the skewness parameter (Cartinhour, 1990) which is a non-linear function of ρ. Thus, they also computed the maximum likelihood estimator of ρ using the proposed likelihood function. However, their approach is limited to the specific case, in which a truncated point of x is $E(x)$.

3.3 Proposed method

We now consider a new approach of parameter estimation of bivariate data (x, y) when the sample is observed with truncated x. We use a marginal distribution and a conditional

distribution to get the joint probability function for (x, y) such that $f(y, x) = f(y \mid x)f(x)$. Since $f(y \mid x, \delta = 1) = f(y \mid x)$, the parameters in the conditional distribution is easy to be obtained. The marginal distribution of x is assumed to be parametrically specified with $f(x; \gamma)$. Parameter γ is obtained by maximizing the observed likelihood that reflects the truncation mechanism.

We assume that the conditional distribution takes the form of a classical linear regression model, given by

$$y = \beta_0 + \beta_1 x + e, \quad e \sim (0, \sigma_e^2), \tag{3.3}$$

where e is uncorrelated with x.

Since $E(e_i \mid x_i, \delta_i = 1) = E(e_i \mid x_i) = 0$ by (3.2), the consistent estimates of $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_e^2)$ can be obtained by the ordinary least squares method. That is, the normality on the error distribution is not needed.

Now, we first consider the case when x is normally distributed with $N(\mu_x, \sigma_x^2)$. To estimate the marginal parameter (μ_x, σ_x^2), note that

$$f(x \mid \delta = 1) = f(x \mid x < x_c) = \frac{\phi\left\{(x_c - \mu_x)/\sigma_x\right\} I(x < x_c)}{\left\{\Phi(x_c - \mu_x)/\sigma_x\right\}},$$

where ϕ is the probability density function and $\Phi(x) = \int_{-\infty}^{x} \phi(z)dz$ is the cumulative density function of standard normal distribution, respectively. The observed log-likelihood is

$$l_2(\mu_x, \sigma_x^2) = -\frac{1}{2} n \log \sigma_x^2 - \frac{1}{2\sigma_x^2} \sum_i (x_i - \mu_x)^2 - \sum_i \log \left\{ \Phi\left(\frac{x_c - \mu_x}{\sigma_x} \right) \right\}. \tag{3.4}$$

Note that the maximum likelihood estimators for (3.4) are the solutions of non-linear equations which are derivatives of $l_2(\mu_x, \sigma_x^2)$ with respect to (μ_x, σ_x^2). One can also use an EM algorithm to obtain the MLE (Kim and Shao, 2013).

We now consider the case when x is not normally distributed. If x is parametrically specified with a density function $f(x; \gamma)$ with an unknown parameter γ, then the observed log-likelihood function is

$$l_2(\gamma) = \sum_i \log f(x_i; \gamma) - \sum_i \log F(x_c; \gamma), \tag{3.5}$$

where $F(x) = \int_{-\infty}^{x} f(x)dx$ is a cumulative density function of x. The maximum likelihood estimator of γ can be obtained by maximizing (3.5).

Once we assume a parametric distribution on truncated x, we can make a goodness-of-fit test to reduce mis-specification risks. For the truncated or censored sample version of Kolmogorov-Smirnov goodness-of-fit tests were studied in various ways such as Barr and Davidson (1973), Koziol and Byar (1975), Dufour and Maag (1978) and Fleming et. al. (1980). Here, we consider modified Kolmogorov-Smirnov test assuming the conditional distribution of truncated x with an unknown parameter γ ,

$$D_n = \sup_{-\infty \le x \le x_c} \mid \hat{F}_n(x) - F_0(x;\gamma) \mid, \tag{3.6}$$

where $\hat{F}_n(x)$ is a empirical cumulative density function with $\hat{F}_n(x) = \frac{1}{n}\sum_i I(x_i \le x)$ and $F_0(x;\gamma)$ is the cumulative distribution function of truncated x which is defined by

$$F_0(x;\gamma) = \int_{-\infty}^{x} \frac{f(t;\gamma)}{F(x_c;\gamma)} dt.$$

Since γ is an unknown parameter, we can use estimated parameters for goodness-of-fit test. Given the estimated parameters, we have a modified test statistic,

$$D_n(\hat{\gamma}) = \sqrt{n} \max_{x_{(i)},\ldots,x_{(n)}} \mid \frac{i}{n} - F_0(x_{(i)};\hat{\gamma}) \mid, \tag{3.7}$$

where x_k is the kth largest value among $x_1, \ldots, x_n$.

Durbin (1975) showed how the modified Kolmogorov-Smirnov test works when parameters are estimated. When the same Kolmogoronov distribution is used to the compute p-value, using the estimated parameters gives more unstable results to specify the distribution of X rather than using true parameters. In practice, one can first consider a normal distribution and perform a goodness-of-fit test. If the test result shows a significant departure from normality, we may consider an alternative model.

We now want to estimate Pearson's correlation coefficient defined by

$$\rho_{xy} = \text{corr}(x,y) = \text{cov}(x,y)/\sqrt{\text{var}(x)\text{var}(y)}.$$

Since the regression slope coefficient β_1 can be expressed with $\text{cov}(x, y)$ and $\text{var}(x)$ under the linear regression,

$$\beta_1 = \text{cov}(x, y)/\text{var}(x),$$

the Pearson's correlation coefficient can be driven in terms of β_1, $\text{var}(x)$ and $\text{var}(y)$ with $\rho = \beta_1\sqrt{\text{var}(x)}/\sqrt{\text{var}(y)}$. Thus, the proposed estimator of ρ would be

$$\hat{\rho} = \hat{\beta}_1\sqrt{\hat{\text{var}}(x)}/\sqrt{\hat{\text{var}}(y)}, \tag{3.8}$$

where $\hat{\beta}_1$ is the maximum likelihood estimator obtained by maximizing of the log-likelihood function of normal distribution and $\hat{\text{var}}(x)$ and $\hat{\text{var}}(y)$ are respectively obtained from the assumed distribution of x and y given x. Since the variance of x could be defined as a function of assumed distribution parameters in general, we can compute $\hat{\text{var}}(x)$ using those parameter estimates. Also, we can compute $\hat{\text{var}}(y)$ with the law of total variance, $\text{var}(y) = E[\text{var}(y \mid x)] + \text{var}[E(y \mid x)]$. For example, we get $\text{var}(y) = \sigma_e^2 + \beta_1^2\text{var}(x)$ in the linear regression model (3.3).

Remark 1 *Instead of assuming a marginal distribution of x, we may assume a conditional distribution of x given y such as another simple linear regression of x on y such that*

$$x = \alpha_0 + \alpha_1 y + u, \quad u \sim N(0, \sigma_u^2). \tag{3.9}$$

Then, another unbiased estimator $\hat{\rho}$ is

$$\hat{\rho} = \hat{\alpha}_1\hat{\beta}_1, \tag{3.10}$$

where $\hat{\alpha}_1$ is obtained by maximizing the following conditional log-likelihood function:

$$\begin{aligned} l_c(\alpha, \sigma_u^2) &= -\frac{1}{2}n\log\sigma_u^2 - \frac{1}{2\sigma_u^2}\sum_i (x_i - \alpha_0 - \alpha_1 y_i)^2 \\ &\quad - \sum_i \log\left\{\Phi\left(\frac{x_c - \alpha_0 - \alpha_1 y_i}{\sigma_u}\right)\right\}. \end{aligned} \tag{3.11}$$

Even if the log-likelihood (3.11) is not globally concave, it has a unique maximum in the interior of parameter space, if one exist (Orme, 1989). Truncated regression has been largely

studied in economics. For example, see Amemiya (1973), Olsen (1980), Goldberger (1981) and Green (1983).

We now discuss variance estimation of the coefficient estimator in (3.8). Note that we can express that

$$\hat{\rho} = \hat{\beta}_1 \left(\frac{\hat{\sigma}_x^2}{\hat{\sigma}_e^2 + \hat{\beta}_1^2 \hat{\sigma}_x^2} \right)^{1/2} \tag{3.12}$$

and the three estimators, $\hat{\beta}_1$, $\hat{\sigma}_e^2$, and $\hat{\sigma}_x^2$, are mutually uncorrelated. We can apply Taylor linearization to obtain linearization variance formula. The closed-form estimates for the variance of $\hat{\rho}$ in (3.12) is

$$\hat{V}(\hat{\rho}) = A^2 \hat{\text{var}}(\hat{\beta}_1) + B^2 \hat{\text{var}}(\hat{\sigma}_x^2) + C^2 \hat{\text{var}}(\hat{\sigma}_e^2), \tag{3.13}$$

where

$$\begin{aligned} A &= \left(\frac{\hat{\sigma}_x^2}{\hat{\sigma}_y^2} \right)^{1/2} \left(1 - \frac{\hat{\sigma}_x^2}{\hat{\sigma}_y^2} \hat{\beta}_1^2 \right), \\ B &= \frac{\hat{\beta}_1}{2\hat{\sigma}_y^2} \left(\frac{\hat{\sigma}_y^2}{\hat{\sigma}_x^2} \right)^{1/2} \left(1 - \frac{\hat{\sigma}_x^2}{\hat{\sigma}_y^2} \hat{\beta}_1^2 \right), \\ C &= -\frac{\hat{\beta}_1}{2\hat{\sigma}_y^2} \left(\frac{\hat{\sigma}_x^2}{\hat{\sigma}_y^2} \right)^{1/2}, \end{aligned}$$

and $\hat{\sigma}_y^2 = \hat{\sigma}_e^2 + \hat{\beta}_1^2 \hat{\sigma}_x^2$. The variance estimates of $\left\{ \hat{\text{var}}(\hat{\beta}_1), \hat{\text{var}}(\hat{\sigma}_e^2) \right\}$ can be directly computed from the inverse of expected Fisher information or observed Fisher information with estimated parameters. Furthermore, the variance estimate $\hat{\text{var}}(\hat{\sigma}_x^2)$ can be also estimated through the negative Hessian matrix of (3.4) or Taylor linearization based on variance estimates of $\text{var}(\hat{\gamma})$ in (3.5). For the bivariate normal distribution example, we have

$$\begin{aligned} \hat{\text{var}}(\hat{\beta}_1) &= \frac{\hat{\sigma}_e^2}{\sum_i (x_i - \bar{x})^2}, \\ \hat{\text{var}}(\hat{\sigma}_e^2) &= \left(\frac{n}{2\hat{\sigma}_e^4} - \frac{\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\hat{\sigma}_e^6} \right)^{-1}, \end{aligned}$$

and $\{\hat{\text{var}}(\hat{\mu}_x), \hat{\text{var}}(\hat{\sigma}_x^2)\}$ are approximated by the diagonal of $I(\hat{\mu}_x, \hat{\sigma}_x^2)^{-1}$,

$$I(\mu_x, \sigma_x^2) = - \begin{bmatrix} I_{11}(\mu_x, \sigma_x^2) & I_{12}(\mu_x, \sigma_x^2) \\ I_{12}(\mu_x, \sigma_x^2) & I_{22}(\mu_x, \sigma_x^2) \end{bmatrix},$$

where

$$
\begin{aligned}
I_{11} &= -\frac{n}{\hat{\sigma}_x^2} + n\left\{\frac{(x_c-\hat{\mu}_x)\hat{f}\hat{F}}{\hat{\sigma}_x^2} + \hat{f}^2\right\}/\hat{F}^2, \\
I_{12} &= -\frac{\sum_i (x_i-\hat{\mu}_x)}{\hat{\sigma}_x^4} + n\left[\left\{\frac{(x_c-\hat{\mu}_x)^2}{2\hat{\sigma}_x^4} - \frac{1}{2\hat{\sigma}_x^2}\right\}\hat{f}\hat{F} + \frac{(x_c-\hat{\mu}_x)}{2\hat{\sigma}_x^2}\hat{f}^2\right]/\hat{F}^2, \\
I_{22} &= \frac{n}{2\hat{\sigma}_x^4} - \frac{\sum_i (x_i-\hat{\mu}_x)^2}{\hat{\sigma}_x^6} - n\left[\left\{\frac{3(x_c-\hat{\mu}_x)}{4\hat{\sigma}_x^4} - \frac{(x_c-\hat{\mu}_x)^3}{4\hat{\sigma}_x^6}\right\}\hat{f}\hat{F} - \frac{(x_c-\hat{\mu}_x)^2}{4\hat{\sigma}_x^4}\hat{f}^2\right]/\hat{F}^2,
\end{aligned}
$$

with $\hat{f} = (2\pi\hat{\sigma}_x^2)^{-1/2}\exp\{-(x_c-\hat{\mu}_x)^2/2\hat{\sigma}_x^2\}$ and $\hat{F} = \int_{-\infty}^{x_c}(2\pi\hat{\sigma}_x^2)^{-1/2}\exp\{-(t-\hat{\mu}_x)^2/2\hat{\sigma}_x^2\}\,dt$.

3.4 Simulation Study

We conducted two simulation studies to check the performance of the proposed method. In the first simulation, we generated bivariate samples (x, y) with size n as the first-phase sample and then dropped sub-samples which correspond to $x_i > 1$. We used two levels of n: $n = 400$ and $n = 800$. In the first simulation, the classical linear regression of y on x is assumed:

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

where $e_i \sim N(0, 0.5)$ with $(\beta_0, \beta_1) = (1, 1)$ and x_i are generated from normal distribution $N(1, 1)$ or gamma distribution Gamma$(1, 1)$.

For the second simulation, given the normality of x, we considered non-normal distributions of y given x such that e follows a t-distribution or a gamma distribution. Through those simulation studies, we checked how our proposed estimator is robust to mis-specification of regression model assuming normality.

We generated 2,000 Monte Carlo samples for each simulation study and computed three estimators of ρ:

1. $\hat{\rho}_{naive}$: a naive estimator which is the standard sample correlation estimator with truncated samples,

$$\hat{\rho}_{naive} = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i (x_i-\bar{x})^2 \sum_i (y_i-\bar{y})^2}}.$$

Table 3.1 Monte Carlo Mean bias and variance of $(\hat{\rho})$ under simulation one

$f(x)$	Size	Estimator	Bias	Variance
$N(1,1)$	n=400	$\hat{\rho}_{naive}$	-0.17	0.0019
		$\hat{\rho}_{skew}$	-0.07	0.0511
		$\hat{\rho}_{new}$	-0.00	0.0024
	n=800	$\hat{\rho}_{naive}$	-0.17	0.0010
		$\hat{\rho}_{skew}$	-0.03	0.0182
		$\hat{\rho}_{new}$	-0.00	0.0011
$Gamma(1,1)$	n=400	$\hat{\rho}_{naive}$	-0.45	0.0030
		$\hat{\rho}_{skew}$	-0.55	0.1335
		$\hat{\rho}_{new}$	-0.05	0.0043
	n=800	$\hat{\rho}_{naive}$	-0.45	0.0014
		$\hat{\rho}_{skew}$	-0.58	0.1000
		$\hat{\rho}_{new}$	-0.04	0.0024

2. $\hat{\rho}_{skew}$: a maximum likelihood estimate of ρ obtained by maximizing the likelihood function of the skew normal distribution (Arnold, Beaver & Meeker, 1993),

$$L(\mu_y, \sigma_y, \rho) = \prod_i \frac{2}{\sigma_y} \phi\left(\frac{y-\mu_y}{\sigma_y}\right) \Phi\left(-\lambda \frac{y-\mu_y}{\sigma_y}\right),$$

where $\lambda = \rho(1-\rho^2)^{-1/2}$. See Azzalini (1985) for details of the skew normal distribution.

3. $\hat{\rho}_{new}$: an estimate obtained using the proposed estimator (3.8).

Note that if a truncation point is not equal to the expected value of x, then we have an identification problem in estimation of $\hat{\rho}_{skew}$. See Arnold, Beaver & Meeker (1993) for details.

Table 3.1 presents the Monte Carlo biases and variances of the three estimators under the first simulation and the result of Table 1 shows that our proposed estimator provides nearly unbiased estimates when the marginal distribution of x is correctly specified. The standard sample correlation estimates seriously underestimate the true correlation for both truncated samples. The maximum likelihood estimates of the skew normal distribution are relatively unbiased for the normal case but they are seriously biased for the non-normal cases. Since the correlation ρ is a function of skewness parameter λ in the skew normal distribution, its estimate

Table 3.2 Monte Carlo Mean bias and variance of $(\hat{\rho})$ under simulation two

$f(e)$	Size	Estimator	Bias	Var
$N(1,1)$	n=400	$\hat{\rho}_{naive}$	-0.17	0.0019
		$\hat{\rho}_{skew}$	-0.07	0.0511
		$\hat{\rho}_{new}$	-0.00	0.0024
	n=800	$\hat{\rho}_{naive}$	-0.17	0.0010
		$\hat{\rho}_{skew}$	-0.03	0.0182
		$\hat{\rho}_{new}$	-0.00	0.0011
$t(10)$	n=400	$\hat{\rho}_{naive}$	-0.19	0.0034
		$\hat{\rho}_{skew}$	-0.18	0.1550
		$\hat{\rho}_{new}$	-0.00	0.0058
	n=800	$\hat{\rho}_{naive}$	-0.19	0.0016
		$\hat{\rho}_{skew}$	-0.14	0.1105
		$\hat{\rho}_{new}$	-0.00	0.0027
$Gamma(1,1)$	n=400	$\hat{\rho}_{naive}$	-0.19	0.0039
		$\hat{\rho}_{skew}$	-1.06	0.2047
		$\hat{\rho}_{new}$	0.00	0.0054
	n=800	$\hat{\rho}_{naive}$	-0.19	0.0019
		$\hat{\rho}_{skew}$	-1.06	0.2062
		$\hat{\rho}_{new}$	-0.00	0.0026

severely depends on the skewness of truncated samples and this makes the estimates of true correlation very unstable. Also, for variance estimation, the absolute value of relative biases of $\hat{V}(\hat{\rho}_{new})$ in (3.13) are less than 3% in all cases which shows that the proposed variance estimator is nearly unbiased. The results are not presented here for brevity.

Table 3.2 presents the Monte Carlo biases and variances of the three estimators under the second simulation. Table 3.2 shows that our proposed estimator assuming the normal error also works well under non-normal error distribution in the linear regression of y on x. For a large sample size $(n = 800)$, the regression slope estimates and its inference are robust to departures from the normal error assumption. That is, we can obtain nearly unbiased estimates of β_1 and var$(y \mid x)$ by constructing of a regression with truncated covariate and untruncated dependent variable for bivariate singly truncated samples. This desirable robustness makes the proposed estimator (3.8) consistent under various model scenarios.

3.5 Application

From 2006 to 2013, the 'World Vision' collected the anthropometric and nutritional data from Community-based Management of Acute Malnutrition in South Sudan. Initially these data are recorded as handwriting on the paper sheets, but they are entered into 15 Excel files by the data entering staffs in World Vision East Africa Regional Office. The aim of Community-based Management of Acute Malnutrition program is to reduce the infant or child death due to famine and the procedure of the program is as following. At first, the nutritional status of children aged less than 5 are screened by measuring the mid-upper arm circumference(MUAC). If the children falls on the criteria of severe acute malnutrition, then they were referred to the community treatment center for treatment. Whenever children visited the nutrition center, the full anthropometric measurements were reassessed and recorded on the patients sheet by the health workers with the treatment center. World Vision South Sudan collected the recorded sheets for years and we were asked to analyze the data statistically to evaluate the current procedures and criteria. The quality control process of the collected data resulted in total 3,488 sets of patients data, but some of them are still subject to missingness in basic anthropometric variables.

The World Health Organization (WHO) and the United Nations Children's Fund (UNICEF) recommend to use a criteria to determine malnutrition: either mid-upper arm circumference (MUAC) $<$ 115mm or weight-for-height Z-score (WHZ) $<$-3. However, MUAC is mostly used to identify the malnourished children who should be included in the malnutrition program, because it is challenging to measure the height and weight together with MUAC. Thus, this leads to generation of singly bivariate observations for MUAC measurement from the field operation.

In case of the World Vision data, children whose MUAC are larger than 115mm are also non-randomly selected to be enrolled in the malnutrition program as measurement errors by the community workers. Thus, we choose children whose MUAC is less than 115mm and

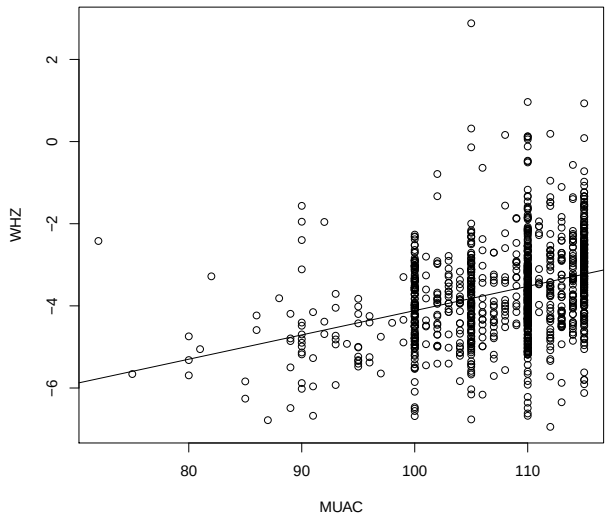

Figure 3.1 Scatter plot with regression line

1,115 cases are refined with both observed MUAC and WHZ.

Since the bivariate samples are singly truncated by MUAC, we firstly assume a linear regression of WHZ on MUAC,

$$\text{WHZ} = \beta_0 + \beta_1 \text{MUAC} + e, \tag{3.14}$$

where $e \sim N(0, 1)$ and MUAC are assumed to be originally generated from truncated normal distribution.

The correlation coefficient estimate is computed by the following four steps:

(Step 1) Estimate the parameters in the specified model.

(Step 2) Conduct a goodness-of-fit test on MUAC with estimated parameters.

(Step 3) Estimate regression slope coefficient β_1 in (3.14) using the ordinary least squares method.

(Step 4) Compute correlation coefficient estimate using (3.8).

Table 3.3 Counts for observed and expected MUAC

	< 105	$105 \sim 115$
Observed	277	838
Expected	290	825

Table 3.4 Regression analysis: MUAC and WHZ

Estimator	Estimates (Standard Error)
$\hat{\beta}_1$	0.059 (0.005)
$\hat{\rho}_{naive}$	0.336 (0.029)
$\hat{\rho}_{new}$	0.743 (0.029)

Since MUAC has many ties at specific points as shown in Figure 3.1, we used the Pearson's chi-squared test instead of the Kolmogorov-Smirnov test. To reduce heaping effect, MUAC is categorized into two bins based on 105 and then we computed observed cases and expected cases. The expected number is calculated using the cumulative density function of conditional distribution with estimated parameters, $\hat{\mu}_{MUAC} = 161$ and $\hat{\sigma}^2_{MUAC} = 426$,

$$n \left\{ \int_{-\infty}^{c_2} f(x; \hat{\mu}_x, \hat{\sigma}_x^2) dx - \int_{-\infty}^{c_1} f(x; \hat{\mu}_x, \hat{\sigma}_x^2) dx \right\} / \int_{-\infty}^{115} f(x; \hat{\mu}_x, \hat{\sigma}_x^2) dx,$$

where x represents MUAC, c_1 and c_2 are determined from $(-\infty, 105, 115)$, and $n = 1,115$. Also the expected number is cumulatively rounded to have integer value. We get the chi-squared test statistic (0.79) and p-value (0.37) from the numbers in Table 3.3. This implies that normal assumption on MUAC is acceptable and it allows us to safely compute correlation coefficient estimate.

Correlation estimation results are summarized in Table 3.4. The proposed method produces a higher correlation coefficient estimate comparing to the naive sample correlation coefficient estimate. The standard error of proposed estimator is calculated using the linearization formula in (3.13).

3.6 Conclusion

Motivated from real data collected by the World Vision, we have considered the Pearson's correction coefficient estimator for a singly truncated bivariate data. A regression model and parametric marginal distribution are assumed instead of bivariate normal distribution on two variables.

In the parametric model for the marginal distribution, one may consider a candidate distribution such as normal distribution and then use a modified Kolmogorov-Smirnov test to test for goodness-of-fit. Once the specified model is accepted, its parameters can be estimated by maximizing the conditional likelihood described in (3.5). For the regression of untruncated variable on truncated variable, we assume a classical linear regression model which allows for non-normal error distributions.

Once we compute the variance of truncated variable and the slope coefficient in the regression model, we can estimate the Pearson's correction coefficient using the simple formula in (3.12). Our simulation study showed that the proposed estimator works well for non-normal truncated variable and also works when the error distribution of linear regression model is no longer normal.

The proposed method is applied to the real malnutrition data collected by World Vision of Africa which measure two malnutrition indicators, WHZ and MUAC. With truncated data, the correlation between two measurements is usually attenuated. The proposed method is applied to the truncated data and the result shows much higher correlation between the two measurements.

4 TWO-PHASE STRATIFIED SAMPLING FOR FRACTIONAL HOT DECK IMPUTATION

Jongho Im[1], Jae kwang Kim[1] and Wayne A. Fuller[1]

Abstract

Hot deck imputation is popular in handling item nonresponse in survey sampling. In hot deck imputation, imputed values are taken from the respondents in the same imputation cell. Imputation cells are used to approximate the imputation model nonparametrically. We extend the fractional hot deck imputation of Kim and Fuller (2004) to the case when some part of imputation cells are also missing. The proposed method of fractional hot deck imputation is performed in two steps and has a similar structure of two-phase stratified sampling. The proposed hot deck imputation method is directly applicable to multivariate missing data with different missing patterns. For variance estimation, we use replication based approach with additional replicates to account for the additional imputation variance. Some numeral results from two simulation studies are also presented.

Key words: Cell mean model; Item nonresponse; EM algorithm, Multivariate missing; Replication variance estimation.

4.1 Introduction

Nonresponse is frequently encountered in survey sampling. Unit nonresponse and item nonresponse are two major types of nonresponse (Kalton and Kasprzyk,1986). While weight-

[1]Department of Statistics, Iowa State University, Ames, U.S.A.

ing adjustment is commonly used to compensate for unit nonresponse, imputation is preferred to handle item nonresponse. Several different imputation methods have been introduced in the literature. Examples of imputation methods include mean imputation, regression imputation, hot deck imputation and nearest neighbor imputation, and so forth. Hazziza (2009) provides a comprehensive overview of those imputation methods.

In household surveys, hot deck imputation is a very popular imputation method. In hot deck imputation, the imputed values are the real observations taken from the respondents in the same sample. Hot deck imputation is popular because it does not create artificial values and also it does not rely on strong model assumptions unlike the imputation method using parametric models. In hot deck imputation, creating imputation cells to achieve homogeneity within imputation cells is critical. In Brick and Kalton (1996), all auxiliary variables are treated as categorical and imputation cells are formed as a combination of those categorized auxiliary variables. A nearest-neighbor imputation approach uses a metric distance of auxiliary variables that is used to find the set of donors (Cotton, 1991; Rancourt, Särndal and Lee, 1994 and Chen and Shao, 2000). Also Hazziza and Beaumont (2007) uses the score estimated by the regression of response on the auxiliary variables or conditional expectation of study variable to create imputation cells.

Variance estimation after hot deck imputation is a challenging problem because it is well known that native approach of treating imputed values as if observed and applying standard variance estimation formula often underestimates the true variance. Rubin (1987) proposed multiple imputation as a general tool for inference with imputed data. In multiple imputation, more than one, say $M(> 1)$, imputed estimates are created for each missing item and then the imputation values are combined using Rubin's formula for variance estimation. Rubin and Schenker (1986) proposed approximate Bayesian bootstrap (ABB) imputation as a hot deck approach to multiple imputation.

On the other hand, instead of multiple imputation, fractional imputation is also proposed (Kalton and Kish, 1984; Kim and Fuller, 2004) as a way of achieving efficient hot deck impu-

tation. Similarly to multiple imputation, M imputed values are generated in fractional imputation, but single data set is created after fractional imputation. Fractional weights are used to handle several imputed values and replication methods are used for variance estimation. Kim and Fuller (2004) and Fuller and Kim (2005) described some properties of fractional hot deck imputation and discussed variance estimation.

In the fractional hot deck imputation of Kim and Fuller (2004), imputation cells are predetermined and the cell mean model is assumed within imputation cells. The determination of imputation cell is not discussed in the Kim and Fuller (2004). In practice, the imputation cells are chosen to achieve homogeneity within imputation cells but sometimes such assumption is not always easy to verify.

In this paper, we consider an extension of fractional hot deck imputation of Kim and Fuller (2004) in two ways. First, instead of assuming that the imputation cells are given, we allow multiple cells for each missing item to account for full uncertainty associated with cell determination. The procedure can be understood as a nonparametric approximation of the true model by a finite mixture model. The implementation of fractional hot deck imputation under the finite mixture model is made through a two-phase stratified sampling mechanism. Second, the proposed method is applied to multivariate missing data with arbitrary missing patterns, using the proposed two-phase stratification approach to determine the imputation cells and compute fractional weights. The joint distribution of the study items are nonparametrically estimated by using a discrete approximation using imputation cells. The joint probabilities of the cells under missing data are computed from a modified EM algorithm and these estimated joint probabilities are used to determine the weights of imputation cells. The replication jackknife variance estimator is proposed for the variance estimation of imputed estimator.

In Section 4.2, the basic setup is introduced. The proposed two-phase stratified fractional imputation and its variance estimation are discussed for a univariate case in Section 4.3. In Section 4.4, the proposed method is extended to the general case of multivariate missing data. Results from two limited simulation studies are presented in Section 4.5. Concluding remarks

are made in Section 4.6.

4.2 Basic setup

Suppose that we have a finite population of size N, indexed by $U = \{1, 2, \cdots, N\}$, and let A be the index set for the units in the sample selected by a probability sampling mechanism. Let A be partitioned into G groups based on the auxiliary information x, where x takes values on $\{1, \cdots, G\}$. Thus, we can write $A = A_1 \cup \cdots \cup A_G$. In addition to x, we collect y and z where y is the study variable and z is another categorical variable that takes values on $\{1, \cdots, H\}$. The cross classification of x and z forms imputation cell and we assume that

$$y_i \mid (x_i = g, z_i = h) \sim ii(\mu_{gh}, \sigma^2_{gh}), \quad i \in U, \tag{4.1}$$

for some μ_{gh} and $\sigma^2_{gh} > 0$, where $\sim ii$ denotes independently and identically distributed. We now write $z_i = (z_{i1}, \ldots, z_{iH})$ and z_{ih} is the indicator function that takes the value one if unit $i \in A_g$ belongs to cell (gh) and is zero otherwise. We assume that x_i is always observed but (y_i, z_i) are subject to missingness. Define $\delta_i = 1$ if (y_i, z_i) is observed and $\delta_i = 0$ otherwise.

We assume that the response mechanism is missing at random (MAR) in the sense that δ is conditionally independent of (y, z) given x. That is,

$$f(y, z \mid x, \delta) = f(y, z \mid x). \tag{4.2}$$

Then, from the conditions (4.2), we have

$$\begin{aligned} f(y \mid x, z, \delta) &= f(y \mid x, z) f(\delta \mid x) / f(\delta \mid x, z) \\ &= f(y \mid x, z), \end{aligned} \tag{4.3}$$

where the second equality comes from condition (4.2). Thus, from result (4.3), then model (4.1) also holds for the responding units. That is,

$$y_i \mid (x_i = g, z_{ih} = 1, \delta_i = 1) \sim ii(\mu_{gh}, \sigma^2_{gh}). \tag{4.4}$$

We now consider a hot deck imputation estimator of $Y_N = \sum_{i=1}^N y_i$ under nonresponse. Since MAR condition (4.2) holds, the following estimator

$$\hat{Y}_I = \sum_{i \in A} w_i \left\{ \delta_i y_i + (1 - \delta_i) E(y_i \mid x_i, \delta_i = 1) \right\}. \tag{4.5}$$

is unbiased for Y_N, where w_i is a sampling weight for unit i.

Now, by the conditions (4.2) and (4.3), we write

$$\begin{aligned} f(y \mid x, \delta = 1) &= f(y \mid x) = \sum_z P(z \mid x) f(y \mid x, z) \\ &= \sum_z P(z \mid x, \delta = 1) f(y \mid x, z, \delta = 1). \end{aligned} \tag{4.6}$$

Model (4.6) takes the form of a finite mixture model. Let $\pi_{h|g} = P(z_h = 1 \mid x = g)$ be the conditional probability of $z_h = 1$ given $x = g$. Here, the variables (x, z) can be understood as the imputation cell variables for hot deck imputation. Note that, from (4.6), we have

$$E(y \mid x = g) = \sum_{h=1}^H \pi_{h|g} E(y \mid x = g, z_h = 1).$$

To construct $E(y_i \mid x_i, \delta_i = 1)$ in (4.5), therefore, we first generate z_i^* from $P(z_i \mid x_i, \delta_i = 1)$ and then generate y_i^* from $f(y_i \mid x_i, z_i^*)$.

Thus, if $\pi_{h|g}$ is known, we can use all the respondents in the cell to estimate $E(y \mid x = g, z_h = 1)$ to get

$$\ddot{Y}_{FEFI} = \sum_{g=1}^G \sum_{i \in A_g} w_i \left\{ \delta_i y_i + (1 - \delta_i) \sum_{h=1}^H \pi_{h|g} \bar{y}_{Rgh} \right\}, \tag{4.7}$$

where

$$\bar{y}_{Rgh} = \frac{\sum_{i \in A_g} w_i \delta_i z_{ih} y_i}{\sum_{i \in A_g} w_i \delta_i z_{ih}}$$

is the weighted mean of respondents in cell (gh). The imputed estimator of (4.7) uses all observed values as donors in the imputation cell and this estimator is often called the fully efficient fractional efficient (FEFI) estimator (Kim and Fuller, 2004). Note that FEFI estimator in (4.7) is unbiased for $Y_N = \sum_{i=1}^N y_i$ under non-informative sampling design and is approximately unbiased under informative sampling design because $E(\bar{y}_{Rgh} \mid x_i = g, z_i = h, \delta_i =$

$1, i \in A_g) \simeq E(y_i \mid x_i = g, z_i = h, \delta_i = 1, i \in U_g)$. In fact, any estimator of the form

$$\ddot{Y}_{FI} = \sum_{g=1}^{G} \sum_{i \in A_g} w_i \left\{ \delta_i y_i + (1 - \delta_i) \sum_{h=1}^{H} \pi_{h|g} \bar{y}^*_{i|gh} \right\} \tag{4.8}$$

is approximately unbiased for Y_N, where $\bar{y}^*_{i|gh} = M_{gh}^{-1} \sum_{k=1}^{M_{gh}} y^{*(k)}_{i|gh}$ and $y^{*(k)}_{i|gh}$ is the k-th imputed value of y_i selected from respondents that belong to the same cell, cell (gh). If $M_{gh} = 2$, we call (4.8) two-per-stratum fractional hot deck imputation which will be further discussed in the following section.

4.3 Fractional hot deck imputation

We now propose a new fractional hot deck imputation (FHDI) procedure that requires less imputation cell information to be known in advance. Given the finite mixture model in (4.6), the imputed values are taken from the imputation cells, where the cell probability for cell (gh) is given by $\pi_{h|g}$. Note that the cell probabilities $\pi_{h|g}$ are unknown and need to be estimated from the sample.

The proposed fractional hot deck imputation is similar in spirit to two-phase sampling for stratification (Rao, 1973; Kim, Navarro, and Fuller, 2006). In phase one, the cells are determined and the cell probabilities $\pi_{h|g}$ are estimated. In phase two, we select M_{gh} donors without replacement in each imputation cell. The detailed procedure is:

[Phase 1]: Estimation of cell probabilities

The $\pi_{h|g}$ are estimated in a nonparametric way so that $\sum_{h=1}^{H} \hat{\pi}_{h|g} = 1$ for each group g. Using the MAR condition, $\pi_{h|g} = Pr(z_{ih} = 1 \mid x_i = g, \delta_i = 1)$, and a nonparametric estimator of $\pi_{h|g}$ is

$$\hat{\pi}_{h|g} = \frac{\sum_{i \in A_g} w_i \delta_i z_{ih}}{\sum_{k \in A_g} w_i \delta_i}, \tag{4.9}$$

which is the estimated relative frequency of z_h for the respondents with $x = g$.

[Phase 2]: Fractional imputation for y within the imputation cell

Once the cell probabilities are estimated, then M_{gh} imputed values are selected from the respondents in each imputation cell (gh) without replacement. Thus, the imputation mechanism can be called two-phase stratified sampling design where the cells are strata. We assume that M_{gh} are no greater than n_{Rgh}, where n_{Rgh} is the number of respondents in cell (gh).

The donors for $y^*_{i|gh}$ are sampled with selection probability proportional to $w^*_{j|gh}$ within the imputation cell, where

$$w^*_{j|gh} = \begin{cases} 0 & \text{if } x_j \neq g, \\ \frac{w_j \delta_j z_{jh}}{\sum_{j \in A_g} w_j \delta_j z_{jh}} & \text{if } x_j = g, \end{cases}$$

so that the imputed values in the cell have equal weights.

Once $\hat{\pi}_{h|g}$ and $\bar{y}^*_{i|gh}$ are obtained, we get the following two-phase fractional imputation (FI) estimator of Y_N based on (4.8), given by

$$\begin{aligned} \hat{Y}_{FI} &= \sum_{g=1}^{G} \sum_{i \in A_g} w_i \left\{ \delta_i y_i + (1 - \delta_i) \sum_{h=1}^{H} \hat{\pi}_{h|g} \bar{y}^*_{i|gh} \right\} \\ &= \sum_{g=1}^{G} \sum_{j \in A_g} w_j \left\{ \delta_j y_j + (1 - \delta_j) \sum_{i \in A_g} \delta_i w^*_{ij} y_i \right\}, \end{aligned} \tag{4.10}$$

where $w^*_{ij} = \sum_{h=1}^{H} \hat{\pi}_{h|g} M^{-1}_{gh} z_{ih} d_{ij}$, where $d_{ij} = 1$ if y_i is used as a donor for y_j and 0 otherwise. Note that $w^*_{ij} \geq 0$ and $\sum_{i \in A_g} w^*_{ij} = 1$.

Note also that

$$\begin{aligned} \sum_{g=1}^{G} \sum_{i \in A_g} w_i \delta_i \sum_{h=1}^{H} \hat{\pi}_{h|g} \bar{y}_{Rgh} &= \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{i \in A_g} w_i \delta_i \frac{\sum_{k \in A_g} w_k \delta_k z_{kh}}{\sum_{k \in A_g} w_k \delta_k} \frac{\sum_{k \in A_g} w_k \delta_k z_{kh} y_k}{\sum_{k \in A_g} w_k \delta_k z_{kh}} \\ &= \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{k \in A_g} w_k \delta_k z_{kh} y_k = \sum_{g=1}^{G} \sum_{k \in A_g} w_k \delta_k y_k. \end{aligned} \tag{4.11}$$

Thus, by (4.11), the fully efficient fractional imputation (FEFI) estimator can be written

$$\begin{aligned}\hat{Y}_{FEFI} &= \sum_{g=1}^{G}\sum_{i\in A_g} w_i\delta_i y_i + \sum_{g=1}^{G}\sum_{i\in A_g} w_i(1-\delta_i)\sum_{h=1}^{H}\hat{\pi}_{h|g}\bar{y}_{Rgh} \\ &= \sum_{g=1}^{G}\sum_{i\in A_g} w_i\delta_i \sum_{h=1}^{H}\hat{\pi}_{h|g}\bar{y}_{Rgh} + \sum_{g=1}^{G}\sum_{i\in A_g} w_i(1-\delta_i)\sum_{h=1}^{H}\hat{\pi}_{h|g}\bar{y}_{Rgh} \\ &= \sum_{g=1}^{G}\sum_{i\in A_g} w_i \sum_{h=1}^{H}\hat{\pi}_{h|g}\bar{y}_{Rgh}. \end{aligned} \tag{4.12}$$

Using the definition of $\hat{\pi}_{h|g}$ and $\bar{y}_{Rgh}$, we can also express (4.12) as

$$\begin{aligned}\hat{Y}_{FEFI} &= \sum_{g=1}^{G}\sum_{i\in A_g} w_i \sum_{h=1}^{H}\hat{\pi}_{h|g}\bar{y}_{Rgh} \\ &= \sum_{g=1}^{G}\hat{R}_g^{-1}\sum_{h=1}^{H}\sum_{i\in A_g} w_i\delta_i z_{ih} y_i \qquad (4.13) \\ &= \sum_{g=1}^{G}\hat{R}_g^{-1}\sum_{i\in A_g} w_i\delta_i y_i, \qquad (4.14)\end{aligned}$$

where $\hat{R}_g = \sum_{i\in A_g} w_i\delta_i / \sum_{i\in A_g} w_i$. Equations (4.13) and (4.14) show that the effect of additional division based on z does not appear in FEFI estimator.

Based on equation (4.14), the estimator $\hat{Y}_{FI}$ can be expressed

$$\hat{Y}_{FI} = \hat{Y}_{FEFI} + \sum_{g=1}^{G}\sum_{i\in A_g} w_i(1-\delta_i)\sum_{h=1}^{H}\hat{\pi}_{h|g}(\bar{y}^*_{i|gh} - \bar{y}_{Rgh}). \tag{4.15}$$

While the fully efficient fractional imputation estimator $\hat{Y}_{FEFI}$ has no variance due to the selection of imputed values, the proposed fractional estimator $\hat{Y}_{FI}$ has additional variance caused by the donor selection procedure. The second part in the right-hand side of (4.15) is subject to the variability due to imputation. The properties of the estimator (4.15) are given in Theorem 1.

Theorem 1 *Consider the following assumptions:*

(A1) A sequence probability sample is drawn from a sequence finite population and $\hat{Y}_n = \sum_{i\in A} w_i y_i$ *is design-unbiased for* Y.

(A2) *The population elements satisfy the cell mean model (4.1) and the MAR condition (4.2) holds for the sequence of populations.*

(A3) *There are at least two observed elements in cell* (gh)*. That is,* $n_{Rgh} \geq 2$*.*

(A4) U_g *is a subset of the finite population that has* $x_i = g$ *with size of* N_g *and*

$$\left(\hat{N}_g - N_g, \hat{N}_{Rg} - N_{Rg}, \hat{Y}_{Rg} - Y_{Rg}\right) = O_p(n^{-1/2}N),$$

where $(\hat{N}_g, \hat{N}_{Rg}, \hat{Y}_{Rg}) = \sum_{i \in A_g} w_i(1, \delta_i, \delta_i y_i)$ *and* $(N_g, N_{Rg}, Y_{Rg}) = \sum_{i \in U_g}(1, \delta_i, \delta_i y_i)$*.*

(A5) $\bar{y}_{Rg} - \mu_g = O_p(n^{-1/2})$*, where* $\bar{y}_{Rg} = Y_{Rg}/N_{Rg}$ *and* $\mu_g = \sum_{h=1}^{H} \pi_{h|g}\mu_{gh}$*.*

The fractional hot deck imputation estimator $\hat{Y}_{FI}$ *in (4.10) is constructed using the two-phase stratified sampling procedure described above with* $M_{gh} \leq n_{Rgh}$ *for all* (gh)*. Then,*

$$\hat{Y}_{FI} = \tilde{Y}_{FI} + o_p(n^{-1/2}N), \tag{4.16}$$

and

$$E(\tilde{Y}_{FI} - Y_N) = 0, \tag{4.17}$$

where

$$\tilde{Y}_{FI} = \tilde{Y}_{FEFI} + \sum_{g=1}^{G}\sum_{i \in A_g} w_i(1 - \delta_i)\sum_{h=1}^{H} \hat{\pi}_{h|g}(\bar{y}^*_{i|gh} - \bar{y}_{Rgh}),$$

and

$$\tilde{Y}_{FEFI} = \sum_{g=1}^{G}\sum_{i \in A_g} w_i\{y_i + (R_g^{-1}\delta_i - 1)(y_i - \mu_g)\},$$

with $R_g = N_{Rg}/N_g$*.*

Also, we have

$$V(\tilde{Y}_{FI}) = V\left(\tilde{Y}_{FEFI}\right) + E\left\{\sum_{g=1}^{G}\sum_{i \in A_g} w_i^2(1 - \delta_i)\sum_{h=1}^{H} \hat{\pi}^2_{h|g}(M_{gh}^{-1} - n_{Rgh}^{-1})\hat{S}^2_{gh}\right\}, \tag{4.18}$$

and

$$V\left(\tilde{Y}_{FEFI}\right) = V\left(\sum_{g=1}^{G}\sum_{i \in A_g} w_i\mu_g\right) + E\left\{\sum_{g=1}^{G} R_g^{-2}\sum_{i \in A_g} w_i^2\delta_i(y_i - \mu_g)^2\right\} \tag{4.19}$$

where $\hat{\pi}_{h|g} = \sum_{i\in A_g} w_i\delta_i z_{ih} / \sum_{k\in A_g} w_i\delta_i$, $\hat{S}^2_{gh} = \hat{v}^{-1}_{gh} \sum_{i\in A_g} w_i\delta_i z_{ih}(y_i - \bar{y}_{Rgh})^2$ *with* $\hat{v}_{gh} = \sum_{i\in A_g} w_i\delta_i z_{ih}$, *and* n_{Rgh} *is the number of respondents in cell* (gh).

See Appendix A for the proof.

In Theorem 1, the proposed FI estimator (4.16) is approximately unbiased and the adjusted FI estimator in (4.17) is unbiased with the cell mean model in (4.1). Unbiasedness of the proposed FI estimator can be directly obtained for non-informative sampling design. Expressions of (4.18) and (4.19) imply that variance of the proposed fractional imputation estimator consists of sampling mechanism, response mechanism and imputation mechanism. Since the second part of right-hand side of (4.18) depends on combinations of imputation cells, the variance of proposed estimator can be varied in different cell combinations. If imputation cells have relatively small values on each σ^2_{gh} and equal size across cells, then the variance of proposed estimator will be almost as efficient as the variance of FEFI estimator.

We now consider variance estimation of $\hat{Y}_{FI}$ using a replication method. To estimate the variance term in (4.17), we use $L+GH$ replicates for variance estimator, where L replicates are used to estimate the first term of (4.17) and the additional GH replicates are used to estimate the second term of (4.17).

Let

$$\hat{V}(\hat{Y}_n) = \sum_{k=1}^{L} c_k(\hat{Y}_n^{(k)} - \hat{Y}_n)^2, \tag{4.20}$$

be a replication variance estimator of $\hat{Y}_n = \sum_{i\in A} w_i y_i$ under complete response, where c_k is the factor associated with the k-th replication and $\hat{Y}_n^{(k)} = \sum_{i\in A} w_i^{(k)} y_i$ is the k-th replicate of $\hat{Y}_n$. We assume that the replication variance estimator (4.20) is consistent for the sampling variance of $\hat{Y}_n$ under complete response.

To discuss estimation of variance in (4.17), the first L replicates, which are created to account for the variance of FEFI estimator, are defined as

$$\hat{Y}^{(k)}_{FI,1} = \sum_{g=1}^{G}\sum_{i\in A_g} w_i^{(k)} \sum_{h=1}^{H} \hat{\pi}^{(k)}_{h|g}\bar{y}^{(k)}_{Rgh} + \sum_{g=1}^{G}\sum_{i\in A_g} w_i(1-\delta_i) \sum_{h=1}^{H} \hat{\pi}_{h|g}(\bar{y}^*_{i|gh} - \bar{y}_{Rgh}), \tag{4.21}$$

where $w_i^{(k)}$ is the k-th replication weight of w_i, $\hat{\pi}_{h|g}^{(k)}$ is the replicate of $\hat{\pi}_{h|g}$ using the k-th replication weights $w_i^{(k)}$, and

$$\bar{y}_{Rgh}^{(k)} = \frac{\sum_{i\in A_g} w_i^{(k)} \delta_i z_{ih} y_i}{\sum_{i\in A_g} w_i^{(k)} \delta_i z_{ih}}.$$

Note that, from the expression (4.21),

$$\hat{Y}_{FI,1}^{(k)} - \hat{Y}_{FI} = \hat{Y}_{FEFI}^{(k)} - \hat{Y}_{FEFI},$$

where $\hat{Y}_{FEFI}^{(k)} = \sum_{g=1}^{G} \{\hat{R}_g^{(k)}\}^{-1} \sum_{i\in A_g} w_i^{(k)} \delta_i y_i$ with $\hat{R}_g^{(k)} = \sum_{i\in A_g} w_i^{(k)} \delta_i / \sum_{i\in A_g} w_i^{(k)}$. Thus, $\sum_{k=1}^{L} c_k (\hat{Y}_{FI,1}^{(k)} - \hat{Y}_{FI})^2$ can be used to estimate the variance term (4.19), the variance of FEFI estimator.

We now want to create replication fractional weights assigned to imputed values in each recipient such that we can express (4.21) as

$$\hat{Y}_{FI,1}^{(k)} = \sum_{g=1}^{G} \sum_{j\in A_g} w_j^{(k)} \{\delta_j y_j + (1-\delta_j) \sum_{i\in A_g} w_{ij}^{*(k)} \delta_i d_{ij} y_i\}, \tag{4.22}$$

where $w_{ij}^{*(k)} \geq 0$ and $\sum_{i\in A_g} w_{ij}^{*(k)} = 1$. Note that the second term of (4.22) uses only the imputed values for each missing unit j, not the whole respondents. To construct a replication estimator in the form of (4.22), we first write expression (4.21) as

$$\begin{aligned} \hat{Y}_{FI,1}^{(k)} = & \sum_{g=1}^{G} \sum_{j\in A_g} w_j^{(k)} \delta_j y_j + \sum_{g=1}^{G} \sum_{j\in A_g} w_j^{(k)} (1-\delta_j) \sum_{h=1}^{H} \hat{\pi}_{h|g}^{(k)} \bar{y}_{Rgh}^{(k)} \\ & + \sum_{g=1}^{G} \sum_{j\in A_g} w_j (1-\delta_j) \sum_{h=1}^{H} \hat{\pi}_{h|g} (\bar{y}_{j|gh}^{*} - \bar{y}_{Rgh}). \end{aligned} \tag{4.23}$$

The second and the third terms of (4.23) cannot be directly expressed as the second term of (4.22).

To express the replication estimates of (4.21) as the imputation form in (4.22), we use Deville and Särndal (1992)'s method. Let replication fractional weights $w_{ij}^{*(k)}$ are obtained by minimizing

$$\sum_{j\in A_g} \sum_{i\in A_g} (w_{ij}^{*(k)} - w_{ij}^{*})^2 \tag{4.24}$$

subject to restrictions

$$\sum_{j\in A_g} w_j^{(k)}(1-\delta_j)\sum_{i\in A_g}\delta_i(w_{ij}^{*(k)}-w_{ij}^{*})y_i \;=\; \hat{Y}_{Mg}^{*}-\hat{Y}_{Mg}^{*(k)}+\hat{N}_{Mg}^{(k)}\bar{y}_{Rg}^{(k)}-\hat{N}_{Mg}\bar{y}_{Rg}, \tag{4.25}$$

where $\hat{Y}_{Mg}^{*}=\sum_{j\in A_g} w_j(1-\delta_j)\bar{y}_{j|g}^{*}$, $\hat{N}_{Mg}=\sum_{j\in A_g} w_j(1-\delta_j)$ and $\bar{y}_{Rg}=\sum_{h=1}^{H}\hat{\pi}_{h|g}\bar{y}_{Rgh}$ and $\hat{Y}_{Mg}^{*(k)}=\sum_{j\in A_g} w_j^{(k)}(1-\delta_j)\bar{y}_{j|g}^{*}$, $\hat{N}_{Mg}^{(k)}=\sum_{j\in A_g} w_j^{(k)}(1-\delta_j)$ and $\bar{y}_{Rg}^{(k)}=\sum_{h=1}^{H}\hat{\pi}_{h|g}^{(k)}\bar{y}_{Rgh}^{(k)}$, where $\hat{\pi}_{h|g}^{(k)}$ and $\bar{y}_{Rgh}^{(k)}$ are obtained by replacing w_j with $w_j^{(k)}$. Then, by the results of Deville and Särndal (1992), an estimator of $w_{ij}^{*(k)}$ which minimizes (4.24) subject to (4.25) can be approximated by a regression estimator. That is, $w_{ij}^{*(k)}$ can be expressed in terms of regression form such that

$$w_{ij}^{*(k)}=w_{ij}^{*}+(\hat{Y}_{Mg}^{*}-\hat{Y}_{Mg}^{*(k)}+\hat{N}_{Mg}^{(k)}\bar{y}_{Rg}^{(k)}-\hat{N}_{Mg}\bar{y}_{Rg})(\Sigma_g^{(k)})^{-1}\delta_i w_{ij}^{*}(y_i-\bar{y}_{j|g}^{*}), \tag{4.26}$$

where $\Sigma_g^{(k)}=\sum_{j\in A_g} w_j^{(k)}(1-\delta_j)\sum_{i\in A_g}\delta_i w_{ij}^{*}(y_i-\bar{y}_{j|g}^{*})^2$ and $\bar{y}_{j|g}^{*}=\sum_{h=1}^{H}\hat{\pi}_{h|g}\bar{y}_{j|gh}^{*}$. From the above construction, we have $\sum_{i\in A_g} w_{ij}^{*(k)}=1$ and

$$\begin{aligned}&\sum_{j\in A_g} w_j^{(k)}(1-\delta_j)\sum_{i\in A_g} w_{ij}^{*(k)}\delta_i y_i\\ =\;&\sum_{j\in A_g} w_j^{(k)}(1-\delta_j)\sum_{h=1}^{H}\hat{\pi}_{h|g}^{(k)}\bar{y}_{Rgh}^{(k)}+\sum_{j\in A_g} w_j(1-\delta_j)\sum_{h=1}^{H}\hat{\pi}_{h|g}(\bar{y}_{j|gh}^{*}-\bar{y}_{Rgh}),\end{aligned}$$

which shows that the replicated fractional weights in (4.26) makes the replicates in (4.22) equal to the replicates in (4.21).

To guarantee non-negativeness of the replication fractional weights in (4.26), we may apply a quadratic programming in the computation of replication fractional weights. The details are presented in Appendix D with an artificial example.

For the second replication estimator, the replicates are created to account for the second term of (4.17) such that

$$E\left\{\sum_{q=1}^{G}\sum_{s=1}^{H}(\hat{Y}_{FI,2}^{(q,s)}-\hat{Y}_{FI})^2\right\}=E\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i^2(1-\delta_i)\sum_{h=1}^{H}\hat{\pi}_{h|g}^2(M_{gh}^{-1}-n_{Rgh}^{-1})\hat{S}_{gh}^2\right\}. \tag{4.27}$$

To estimate the second replicates, assume that M_{gh} be an even number and M_{gh} imputed values are randomly and equally distributed to the first donor group and second donor group. Then, the second GH $(q = 1 \ldots, G; s = 1, \ldots, H)$ replicates are computed by algebraically,

$$\hat{Y}_{FI,2}^{(q,s)} = \sum_{g=1}^{G} \sum_{i \in A_g} w_i \left\{ \delta_i y_i + (1 - \delta_i) \sum_{h=1}^{H} \hat{\pi}_{h|g} \bar{y}_{i|gh}^{*(q,s)} \right\}. \tag{4.28}$$

where

$$\bar{y}_{i|gh}^{*(q,s)} = \bar{y}_{i|gh}^{*} + \phi_{gh}^{(q,s)} \left(\sum_{j \in D_{i1}} y_{i|gh}^{*(i),1} - \sum_{j \in D_{i2}} y_{i|gh}^{*(i),2} \right) \zeta_i^{(q,s)},$$

D_{il} is a set of l-th group of imputed values for unit i, $\zeta_i^{(q,s)}$ is an independent variable taking 1 or -1 independently with equal probability and

$$\phi_{gh}^{(q,s)} = \begin{cases} \phi_{gh} & \text{if } q = g, s = h \\ 0 & \text{otherwise,} \end{cases}$$

and ϕ_{gh} are to be determined to satisfy (4.27). Here, $y_{i|gh}^{*(j),1}$ and $y_{i|gh}^{*(j),2}$ are respectively the j-th imputed value of the first group and the second group in cell (gh).

Now, we consider

$$\begin{aligned} & E\left\{ \sum_{q=1}^{G} \sum_{s=1}^{H} (\hat{Y}_{FI,2}^{(q,s)} - \hat{Y}_{FI})^2 \right\} \\ = \; & E\left\{ \sum_{g=1}^{G} \sum_{i \in A_g} w_i^2 (1 - \delta_i) \sum_{h=1}^{H} \hat{\pi}_{h|g}^2 \phi_{gh}^{2(q,s)} \left(\sum_{j \in D_{i1}} y_{i|gh}^{*(i),1} - \sum_{j \in D_{i2}} y_{i|gh}^{*(i),2} \right)^2 \right\} \\ = \; & E\left\{ \sum_{g=1}^{G} \sum_{i \in A_g} w_i^2 (1 - \delta_i) \sum_{h=1}^{H} \hat{\pi}_{h|g}^2 \phi_{gh}^{2(q,s)} V_I \left(\sum_{j \in D_{i1}} y_{i|gh}^{*(i),1} - \sum_{j \in D_{i2}} y_{i|gh}^{*(i),2} \right) \right\} \\ = \; & E\left\{ \sum_{g=1}^{G} \sum_{i \in A_g} w_i^2 (1 - \delta_i) \sum_{h=1}^{H} \hat{\pi}_{h|g}^2 \phi_{gh}^{2(q,s)} M_{gh} \hat{f}_{Rgh} \hat{S}_{gh}^2 \right\}, \end{aligned} \tag{4.29}$$

where $\hat{f}_{Rgh} = n_{Rgh}/(n_{Rgh} - 1)$, the last equality holds from $\mathrm{Cov}_I\{y_{i|gh}^{*(1)}, y_{i|gh}^{*(2)}\} = -(n_{Rgh} - 1)^{-1} \hat{S}_{gh}^2$ and $V_I(\cdot)$ and $\mathrm{Cov}_I(\cdot, \cdot)$ are variance and covariance with respect to the imputation mechanism, respectively.

If ϕ_{gh} is determined to satisfy $M_{gh}\hat{f}_{gh}\phi_{gh}^2 = \left(M_{gh}^{-1} - n_{Rgh}^{-1}\right)$, then

$$\sum_{g=1}^{G}\sum_{i\in A_g} w_i^2(1-\delta_i)\sum_{h=1}^{H}\hat{\pi}_{h|g}^2 M_{gh}\hat{f}_{gh}\phi_{gh}^2 = \sum_{g=1}^{G}\sum_{i\in A_g} w_i^2(1-\delta_i)\sum_{h=1}^{H}\hat{\pi}_{h|g}^2(M_{gh}^{-1} - n_{Rgh}^{-1})$$

and (4.27) is satisfied. Thus, the replicates in (4.28) can be expressed as

$$\hat{Y}_{FI,2}^{(q,s)} = \sum_{g=1}^{G}\sum_{j\in A_g} w_j\left\{\delta_j y_j + (1-\delta_j)\sum_{i\in A_g}\delta_i(w_{ij}^{*} + w_{ij}^{*(q,s)})d_{ij}y_i\right\},$$

where $w_{ij}^{*(q,s)} = \sum_{h=1}^{H}\hat{\pi}_{h|g}\phi_{gh}^{(q,s)}\zeta_i^{(q,s)}q_{ij,gh}$ with,

$$q_{ij,gh} = \begin{cases} 1 & \text{for the first donor group in cell } (gh), \\ -1 & \text{for the second donor group in cell } (gh), \\ 0 & \text{others.} \end{cases}$$

We have that $w_{ij}^{*} + w_{ij}^{*(q,s)} \geq 0$ because $\phi_{gh} = \{M_{gh}^{-1}\hat{f}_{gh}^{-1}(M_{gh}^{-1} - n_{Rgh}^{-1})\}^{1/2} \leq M_{gh}^{-1}$ and $\sum_{i\in A_g}(w_{ij}^{*} + w_{ij}^{*(q,s)}) = 1$ for all q and s because the sizes of D_{i1} and D_{i2} are equal to each other.

If $M_{gh} > 2$ is odd, we first assign zero value to a randomly selected imputed value and then $M_{gh} - 1$ imputed values are randomly and divided into two equal sized groups. Also, $\zeta_i^{(q,s)}$ independently has $\{M_{gh}/(M_{gh}-1)\}^{1/2}$ or $-\{M_{gh}/(M_{gh}-1)\}^{1/2}$ with equal probability. Then, we have

$$V_I\left\{\phi_{gh}^{(q,s)}\left(\sum_{j\in D_{i1}} y_{i|gh}^{*(i),1} - \sum_{j\in D_{i2}} y_{i|gh}^{*(i),2}\right)\zeta_i^{(q,s)}\right\}^2 = \phi_{gh}^{2(q,s)}M_{gh}\hat{f}_{Rgh}$$

and $w_{ij}^{*} + w_{ij}^{*(q,s)} \geq 0$ and $\sum_{i\in A_g}(w_{ij}^{*} + w_{ij}^{*(q,s)}) = 1$ are guaranteed by ϕ_{gh} that is estimated to satisfy $M_{gh}\hat{f}_{Rgh}\phi_{gh}^2 = \left(M_{gh}^{-1} - n_{Rgh}^{-1}\right)$.

Theorem 2 provides an approximately unbiased replication variance estimator of $\hat{Y}_{FI}$ using two types of replicates discussed in the above. The proposed variance estimator in Theorem 2 consists of the two terms to account for the total variance of FI estimator in (4.18). The first term is constructed by the replicates in (4.21) and the second consists of the replicates in (4.28).

Theorem 2 *In addition to* $(A1)$-$(A4)$ *in Theorem 1, assume further:*

(A6) Let $\hat{V}(\hat{\gamma}_n)$ *be the complete-sample replicate estimator of the variance of a total,* $\hat{\gamma}_n = \sum_{i \in A} w_i \gamma_i$*, and assume that, for any* γ *with bounded fourth moments,*

$$E\left[\left\{V(\hat{\gamma}_n \mid \mathcal{F}_N)^{-1}\hat{V}(\hat{\gamma}_n) - 1\right\}^2 \mid \mathcal{F}_N\right] = o(1),$$

where $\mathcal{F}_N$ *is a set of finite population.*

(A7) ${c_k}^{1/2}\left(\hat{N}_g^{(k)} - \hat{N}_g, \hat{N}_{Rg}^{(k)} - \hat{N}_{Rg}, \hat{Y}_{Rg}^{(k)} - \hat{Y}_{Rg}\right) = O_p(n^{-1}N)$
where $(\hat{N}_g^{(k)}, \hat{N}_{Rg}^{(k)}, \hat{Y}_{Rg}^{(k)}) = \sum_{i \in A_g} w_i^{(k)}(1, \delta_i, \delta_i y_i)$ *are the* k*-th replicates of* $(\hat{N}_g, \hat{N}_{Rg}, \hat{Y}_{Rg}) = \sum_{i \in A_g} w_i(1, \delta_i, \delta_i y_i)$.

(A8) $N^{-1}\sum_{i=1}^{N} y_i^{2+\tau} = O(1)$ *for some* $\tau \geq 2$.

The proposed replication estimator using replicates (4.21) and (4.28) is given by

$$\hat{V}\left(\hat{Y}_{FI}\right) = \sum_{k=1}^{L} c_k(\hat{Y}_{FI,1}^{(k)} - \hat{Y}_{FI})^2 + \sum_{q=1}^{G}\sum_{s=1}^{H}(\hat{Y}_{FI,2}^{(q,s)} - \hat{Y}_{FI})^2, \tag{4.30}$$

where c_k *is the factor associated with the* k*-th replication of* $\hat{Y}_{FI,1}$*,* $\hat{Y}_{FI,1}^{(k)}$ *is defined in (4.21) and* $\hat{Y}_{FI,2}^{(k)}$ *is defined in (4.28) with* $\hat{\phi}_{gh} = \{M_{gh}^{-1}\hat{f}_{Rgh}^{-1}(M_{gh}^{-1} - n_{Rgh}^{-1})\}^{1/2}$*. Then, the replication variance estimator satisfies*

$$\hat{V}\left(\hat{Y}_{FI}\right) = V\left(\tilde{Y}_{FI}\right) - \sum_{g=1}^{G}\sum_{i \in U_g} V(R_g^{-1}\delta_i)\sigma_g^2 + o_p(n^{-1}N^2), \tag{4.31}$$

where U_g *is a subset of finite population that has* $x_i = g$ *with size of* N_g.

See Appendix B for the proof.

If δ_i follows a Bernoulli distribution within each group g with $E(\delta_i) = R_g$, then the second term of (4.31) becomes $\sum_{g=1}^{G}\sum_{i \in U_g} R_g^{-1}(1 - R_g)\sigma_g^2$, which is also presented in Kim, Navarro and Fuller (2006). Unless the sample size n is large with respect to N, this term can be ignored. If the sample size is large relative to N and δ_i is assumed to follow a Bernoulli distribution, then an estimator,

$$\sum_{g=1}^{G}\hat{R}_g^{-2}(1 - \hat{R}_g)\sum_{i \in A_g} w_i\delta_i(y_i - \bar{y}_{Rg})^2$$

can be incorporated in $\hat{V}(\hat{Y}_{FI})$.

4.4 Extension to Multivariate missing data

Creating hot deck imputation for multivariate missing data is a notoriously challenging problem. Judkins et al. (2007) proposed an iterative hot deck imputation procedure similar to Gibbs sampler in the sense that covariance structures are preserved through iterations. They did not give a variance estimation. Shao and Wang (2002) proposed a joint regression imputation that preserves marginal totals, second moments and correlation structure of bivariate survey data. Shao and Wang (2002) considered a jackknife method for variance estimation.

We now extend the proposed method in Section 3 to multivariate missing case, $\mathbf{y} = (y_1, \cdots, y_p)$. Let δ_{ik} be the response indicator function for y_{ik}. For each item k, assume that we have discretized values of y_k, denoted by $\tilde{y}_k$, based on the sample quantiles in the respondents set or $\tilde{y}_k$ can be predetermined such that the cell mean model holds within the cell determined by $(\tilde{y}_1, \ldots, \tilde{y}_p)$. If $\tilde{y}_k$ takes G_k distinct values for item k, then the total number of cells for p variables is $G_T = G_1 \times \cdots \times G_p$. The cross-classification of $\tilde{y}$ defines imputation cells in multivariate missing data. Given the realized responses, we can write $\mathbf{y}_i = (\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis})$, where $\mathbf{y}_{i,obs}$ and $\mathbf{y}_{i,mis}$ are the observed part and missing part of $\mathbf{y}_i$, respectively. Similarly, we can write $\tilde{\mathbf{y}}_i = (\tilde{\mathbf{y}}_{i,obs}, \tilde{\mathbf{y}}_{i,mis})$. Strictly speaking, the partition $(\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis})$ can be different for each i. Thus, if is understood that $obs = obs(i)$ and $mis = mis(i)$.

Table 4.1 illustrates the data format obtained after discretization. Three study variables y_1, y_2 and y_3 are respectively categorized into two groups. The first eight units are fully observed, that is, $\mathbf{y}_{i,obs} = (y_{1i}, y_{2i}, y_{3i})$ and $\mathbf{y}_{i,mis}$ are empty vectors. However, other units have missing values. If study variable y is missing, then discretized value $\tilde{y}$ is also missing.

We now decompose an index set A into two subsets A_R and A_M such that $A = A_R \cup A_M$, where $A_R (\subset A)$ contains indexes of units that are are fully observed for all items and $A_M (\subset A)$ contains indexes of units that have at least one missing item. Thus, if a unit belongs to A_R, then $\mathbf{y}_i = \mathbf{y}_{i,obs}$ and $\tilde{\mathbf{y}}_i = \tilde{\mathbf{y}}_{i,obs}$. In the illustrative example in Table 4.1, $A_R = \{1, 2, 3, 4, 5, 6, 7, 8\}$ and $A_M = \{9, 10\}$.

Table 4.1 An illustrative example with multivariate missing data ($p = 3$)

A	ID	Weight	y_1	y_2	y_3	$\tilde{y}_1$	$\tilde{y}_2$	$\tilde{y}_3$
A_R	1	w_1	$y_{1,1}$	$y_{2,1}$	$y_{3,1}$	1	1	2
	2	w_2	$y_{1,2}$	$y_{2,2}$	$y_{3,2}$	1	1	2
	3	w_3	$y_{1,3}$	$y_{2,3}$	$y_{3,3}$	2	1	1
	4	w_4	$y_{1,4}$	$y_{2,4}$	$y_{3,4}$	2	1	1
	5	w_5	$y_{1,5}$	$y_{2,5}$	$y_{3,5}$	2	1	1
	6	w_6	$y_{1,6}$	$y_{2,6}$	$y_{3,6}$	2	2	2
	7	w_7	$y_{1,7}$	$y_{2,7}$	$y_{3,7}$	2	2	2
	8	w_8	$y_{1,8}$	$y_{2,8}$	$y_{3,8}$	2	2	2
A_M	9	w_9	?	$y_{2,9}$	$y_{3,9}$	2	?	2
	10	w_{10}	?	$y_{2,10}$	?	?	1	?

?: missing value

A proposed strategy is that missing items of a unit in A_M are jointly imputed from the units in A_R. For a donor unit, we take all values of the donor corresponding to missing items of recipient. For example, if unit 1 is selected as a donor for unit 10, $(y_{1,1}, y_{3,1})$ is the imputed vector for $(y_{1,10}, y_{3,10})$ leading to $\mathbf{y}^*_{10} = (y^*_{1,10}, y_{2,10}, y^*_{3,10}) = (y_{1,1}, y_{2,10}, y_{3,1})$.

Now, we discuss how imputation cells are created and donors are selected for recipients based on discretized values $\tilde{\mathbf{y}}$. First, for recipient $i \in A_M$, find all possible discretized values for missing items from respondents in A_R who have the same cell values as $\tilde{\mathbf{y}}_{i,obs}$. Then, a set of possible values of $\tilde{\mathbf{y}}_{mis(i)}$ is defined by $\mathcal{H}_i = \{\tilde{\mathbf{y}}_{j,mis(i)}; \tilde{\mathbf{y}}_{j,obs(i)} = \tilde{\mathbf{y}}_{i,obs(i)}, \delta_j = 1\}$, where $\delta_j = \prod_{k=1}^{p} \delta_{jk}$ and $\tilde{\mathbf{y}}_{j,obs(i)}$ and $\tilde{\mathbf{y}}_{j,mis(i)}$ are partial vector of $\tilde{\mathbf{y}}_j$ that correspond to observed part and missing part of $\tilde{\mathbf{y}}_i$, respectively. We now let H_i be the size of $\mathcal{H}_i$, where all H_i elements of $\mathcal{H}_i$ can be used as a donor for unit i. Thus, $\mathcal{A}_{ih} = \{j; \tilde{\mathbf{y}}_{j,mis(i)} = \tilde{\mathbf{y}}_{h,mis(i)}, \tilde{\mathbf{y}}_{h,mis(i)} \in \mathcal{H}_i\}$ is the set of possible donors corresponding to the h-th imputation cell for recipient i, $h = 1 \cdots, H_i$. After constructing $\mathcal{H}_i$ and $\mathcal{A}_{ih}$, we select M_h donors from $\mathcal{A}_{ih}$ for the h-th imputation cell. Thus, $(\tilde{\mathbf{y}}_{i,obs(i)}, \tilde{\mathbf{y}}_{i,mis(i)})$ plays the role of x_i and z_i in Section 4.3, respectively.

In the example in Table 4.1, for unit 9, we have a single index set, $\mathcal{A}_{9,1} = \{6, 7, 8\}$, because all possible donors have the same value of $\tilde{y}_2 = 2$. With similar way, we have $\mathcal{A}_{10,1} = \{1, 2\}$, $\mathcal{A}_{10,2} = \{3, 4, 5\}$ and $\mathcal{H}_{10} = \{(1, 2), (2, 1)\}$ for imputation of $(\tilde{y}_{1,10}, \tilde{y}_{3,10})$.

To implement the proposed hot deck imputation, a necessary condition is that $\mathcal{H}_i$ should be non-empty for all $i \in A_M$. Also, there should be at least two fully observed respondents whose $\tilde{\mathbf{y}}_{mis(i)}$ correspond to an element of $\mathcal{H}_i$. Unless these two properties are met, we need a cell collapsing strategy. An illustration of the cell collapsing strategy is introduced in Appendix E.

Now, similarly to (4.6), the conditional distribution of $f(\mathbf{y}_{i,mis} \mid \mathbf{y}_{i,obs})$ can be approximated by

$$f(\mathbf{y}_{i,mis} \mid \mathbf{y}_{i,obs}) \cong \sum_{\tilde{\mathbf{y}}^*_{i,mis} \in \mathcal{H}_i} p(\tilde{\mathbf{y}}^*_{i,mis} \mid \tilde{\mathbf{y}}_{i,obs}) f(\mathbf{y}_{i,mis} \mid \tilde{\mathbf{y}}_{i,obs}, \tilde{\mathbf{y}}^*_{i,mis}), \tag{4.32}$$

where $p(\tilde{\mathbf{y}}^*_{i,mis} \mid \tilde{\mathbf{y}}_{i,obs})$ is the conditional cell probability of $\tilde{\mathbf{y}}^*_{i,mis}$ given $\tilde{\mathbf{y}}_{i,obs}$. Since the expression (4.32) have the same mixture model structure of univariatae case in (4.6), under MAR assumption, we can mimic the two-phase stratified sampling method introduced in Section 3 as follows:

[Phase 1]: Estimation of cell probabilities

First step for the multivariate hot deck imputation is to assign all possible H_i imputation cell vectors on each recipient with conditional cell probabilities. To compute $p(\tilde{\mathbf{y}}^*_{i,mis} \mid \tilde{\mathbf{y}}_{i,obs})$, we first need to estimate $p(\tilde{\mathbf{y}})$, where $p(\tilde{\mathbf{y}})$ is a cell probability of obtaining a particular value of $\tilde{\mathbf{y}}$. Since we have missing items on $\tilde{\mathbf{y}}_{i,mis}$, we cannot directly estimate cell probabilities as in the univariate missing case. From the partial observations, we can use a modified EM algorithm to estimate the joint probabilities, $\hat{p}(\tilde{\mathbf{y}})$. The procedure avoids producing positive probabilities for structural zeros. See Appendix C for a description of the modified EM algorithm. For an imputed cell vector $\tilde{\mathbf{y}}^*_{i,mis} \in \mathcal{H}_i$, the estimated conditional cell probability $\hat{p}(\tilde{\mathbf{y}}^*_{i,mis} \mid \tilde{\mathbf{y}}_{i,obs})$ can be computed using

$$\hat{p}(\tilde{\mathbf{y}}^*_{i,mis} \mid \tilde{\mathbf{y}}_{i,obs}) = \hat{p}(\tilde{\mathbf{y}}^*_i) / \sum_{h=1}^{H_i} \hat{p}(\tilde{\mathbf{y}}_i^{*(h)}), \tag{4.33}$$

where $\tilde{\mathbf{y}}_i^{*(h)} = (\tilde{\mathbf{y}}_{i,obs}, \tilde{\mathbf{y}}^{*(h)}_{i,mis})$ and $\tilde{\mathbf{y}}^{*(h)}_{i,mis}$ is the h-th imputation cell vector that belongs $\mathcal{H}_i$.

[Phase 2]: Fractional imputation for $\mathbf{y}_{mis}$ within the imputation cell

The fractional hot deck imputation for multivariate case can be implemented with the same process of univariate case. In h-th imputation cell corresponding $\tilde{\mathbf{y}}_{i,mis}^{*(h)}$, $M_h(=2)$ imputed values are sampled without replacement. The j-th donor for the unit i in the h-th imputation cell is sampled with the selection probability proportional to $w_{j|ih}^*$, where

$$w_{j|ih}^* = w_j \gamma_{ijh} / \sum_{j \in A_R} w_j \gamma_{ijh}$$

and γ_{ijh} is a matching indicator function that have the value one if $\tilde{\mathbf{y}}_{j,mis(i)} = \tilde{\mathbf{y}}_{i,mis}^{*(h)}$ and $\tilde{\mathbf{y}}_{j,mis(i)} \in \mathcal{H}_i$ and zero otherwise. We assume that the size of candidates $R_{ih} = \sum_{j \in A_R} \gamma_{ijh}$ is greater than M_h. Note that we have $M_i = H_i M_h$ total imputed values for recipient i.

Table 4.2 Final imputed data for the illustrative example

A	ID	Weight	y_1	y_2	y_3	$\tilde{y}_1$	$\tilde{y}_2$	$\tilde{y}_3$
A_M	9	$w_9 w_{9,1}^*$	$y_{1,9}$	$y_{2,7}^*$	$y_{3,9}$	2	2^*	2
		$w_9 w_{9,2}^*$	$y_{1,9}$	$y_{2,8}^*$	$y_{3,9}$	2	2^*	2
	10	$w_{10} w_{10,1}^*$	$y_{1,1}^*$	$y_{2,10}$	$y_{3,1}^*$	1^*	1	2^*
		$w_{10} w_{10,2}^*$	$y_{1,2}^*$	$y_{2,10}$	$y_{3,2}^*$	1^*	1	2^*
		$w_{10} w_{10,3}^*$	$y_{1,3}^*$	$y_{2,10}$	$y_{3,3}^*$	2^*	1	1^*
		$w_{10} w_{10,4}^*$	$y_{1,5}^*$	$y_{2,10}$	$y_{3,5}^*$	2^*	1	1^*

y^*: imputed values

Table 4.2 shows one realization of the results from the proposed multivariate hot deck imputation for the illustrative example in Table 4.1. For unit 9, unit 7 and unit 8 are selected from $\mathcal{A}_{9,1}$ and $y_{2,7}^*$ and $y_{2,8}^*$ are imputed in position of $y_{2,9}$. For tenth unit, there are two possible imputation cell vectors, $\{(1,2),(2,1)\}$ for $(\tilde{y}_{1,10}, \tilde{y}_{3,10})$. The first and second units are drawn from $\mathcal{A}_{10,1}$ and the third and fifth units are selected from $\mathcal{A}_{10,2}$ with the selection probability proportional to the sampling weights. Values of selected donors are jointly imputed in position of $m_{1,10}$ and $m_{3,10}$. Fractional weights for imputation cells are computed by using (4.33). Fractional weights for the imputed value for unit 9 are $w_{9,1}^* = w_{9,2}^* = M_h^{-1} \hat{p}(\tilde{y}_{2,9}^* = 2 \mid \tilde{y}_{1,9} = 2, \tilde{y}_{3,9} = 2)$. With similar way, $w_{10,1}^* = w_{10,2}^* = M_h^{-1} \hat{p}(\tilde{y}_{1,10}^* = 1, \tilde{y}_{3,10}^* = 2 \mid \tilde{y}_{2,10} = 1)$ and

$w^*_{10,3} = w^*_{10,4} = M_h^{-1}\hat{p}(\tilde{y}^*_{1,10} = 2, \tilde{y}^*_{3,10} = 1 \mid \tilde{y}_{2,10} = 1)$. Here, all fractional weights are nonnegative and summation over each recipient equals to one.

Once we have imputed values and fractional weights after the above second phase, we get the FI estimator of $Y_l = \sum_{i=1}^N y_{ik}$, $(l = 1, \ldots, p)$ with the same expression of (4.10),

$$\begin{aligned}\hat{Y}_{l,FI} &= \sum_{j\in A} w_j \left\{ \delta_{jl} y_{jl} + (1-\delta_{jl}) \sum_{h=1}^{H_j} \hat{p}(\tilde{\mathbf{y}}^{*(h)}_{j,mis} \mid \tilde{\mathbf{y}}_{j,obs}) \bar{y}^*_{j|h} \right\}, \\ &= \sum_{j\in A} w_j \left\{ \delta_{jl} y_{jl} + (1-\delta_{jl}) \sum_{i\in A_R} \delta_i w^*_{ij} y_{il} \right\}\end{aligned}$$

where $w^*_{ij} = \sum_{h=1}^{H_j} \hat{p}(\tilde{\mathbf{y}}^{*(h)}_{j,mis} \mid \tilde{\mathbf{y}}_{j,obs}) M_h^{-1} \gamma_{jih} d_{ij}$ and $\bar{y}^*_{j|h}$ is mean of imputed values in the h-th imputation cell for unit j. Note that $w^*_{ij} \geq 0$ and $\sum_{i\in A_R} w^*_{ij} = 1$.

For variance estimation, we write $\hat{Y}_{l,FI}$ again such that

$$\hat{Y}_{l,FI} = \hat{Y}_{l,FEFI} + \sum_{j\in A} w_j (1-\delta_{jl}) \sum_{h=1}^{H_j} \hat{p}(\tilde{\mathbf{y}}^{*(h)}_{j,mis} \mid \tilde{\mathbf{y}}_{j,obs}) (\bar{y}^*_{j|h} - \bar{y}^*_{j|Rh}), \tag{4.34}$$

where $\bar{y}^*_{j|Rh}$ is the mean of all observed values in the h-th imputation cell corresponding to unit j and

$$\hat{Y}_{l,FEFI} = \sum_{j\in A} w_j \left\{ \delta_{jl} y_{jl} + (1-\delta_{jl}) \sum_{h=1}^{H_j} \hat{p}(\tilde{\mathbf{y}}^{*(h)}_{j,mis} \mid \tilde{\mathbf{y}}_{j,obs}) \bar{y}^*_{j|Rh} \right\}.$$

Since expression (4.34) have the same form with (4.15) , we can also use the same replication estimator in (4.21) and (4.22). Also, for the second type replicates, we can use (4.28) and ϕ_{ih} corresponding to ϕ_{gh} in Section 3 can be obtained from by solving $M_h \hat{f}_{Rih} \phi^2_{ih} = (M_h^{-1} - R_{ih}^{-1})$ in each imputation cell for unit i, where $\hat{f}_{Rih} = n_{Rih}/(n_{Rih} - 1)$ and n_{Rih} is the size of all possible candidates in A_R corresponding imputation cell h for unit i.

4.5 Simulation Study

4.5.1 Univariate missing case

To check the performance of the proposed method in the univariate case, we performed two simulation studies. In the first simulation, an infinite population is assumed and $Y_i = (Y_{1i}, Y_{2i})$,

$i = 1, \cdots, n$ are randomly generated from

$$\begin{aligned} Y_1 &\sim U(0,2), \\ Y_2 &= 1 + Y_1 + e_2, \end{aligned}$$

where e_2 is independent of Y_1 and is generated from a normal distribution, $N(0, 1/2)$. Here, Y_1 is fully observed observed but Y_2 is subject missingness with the response 0.7 for all sample. Thus, Y_1 plays the role of x in the method of Section 3. We used two sample sizes, $n = 300$ and $n = 500$.

To implement the fractional hot deck imputation, Y_1 and Y_2 are categorized into $\tilde{Y}_1$ and $\tilde{Y}_2$ that respectively play roles of x and z of Section 2. The auxiliary variable ,Y_1, is categorized into five groups and the study variable, Y_2, is partitioned into two groups based on the sample quantiles of the respondents. For example, observations with y_2 values less than the median belong to group 1 (i.e. $\tilde{y}_2 = 1$). To implement the proposed fractional hot deck imputation, we impute two values of $\tilde{y}_2$, $\tilde{y}_2 = 1$ and $\tilde{y}_2 = 2$, and $M_{gh} = 2$ imputed values are taken from the respondents in the cell. The imputed values are assigned with the cell fractional weights $\hat{\pi}_{h|g} = \hat{P}(\tilde{y}_2 = h \mid \tilde{y}_1 = g)$. Since we have two groups for $\tilde{y}_2$, there are $m = 4$ imputed values for each recipient. We used $B = 2,000$ Monte Carlo samples in this simulation.

We consider five parameters: $\theta_1 = E(Y_2)$, $\theta_2 = P(Y_2 < 2)$, $\theta_3 = E(Y_2 \mid D = 1)$ with $D \sim Bernoulli(0.3)$, θ_4 is the slope of regression of Y_2 on Y_1 and θ_5 is the correlation between Y_1 and Y_2. These parameters are estimated with three different methods:(1) Full: full sample estimator using n elements (ii) FEFI: fully efficient fractional estimator using (4.12), and (iii) FI: the proposed estimator using (4.10) with $m = 4$. The full sample estimator of $(\theta_1, \theta_2, \theta_3)$ is

$$\hat{\theta}_{\text{Full}} = \sum_{i \in A} z_i / \sum_{i \in A} s_i,$$

where $z_i = \{y_{2i}, I(y_{2i} < 2), d_i y_{2i}\}$, $s_i = \{1, 1, d_i\}$ and d_i is an indicator function with $d_i = 1$ if i is in the domain and 0 otherwise. For the fractional imputation estimator, the numerator $\sum_{i \in A} z_i$ is replaced by $\sum_{j \in A} \delta_{2j} z_j + \sum_{j \in A} (1 - \delta_j) \sum_{i \in A} \delta_{2i} w_{ij}^* z_i$, w_{ij}^* is the i-th fractional

weight for the recipient j obtained in Section 3. The imputed estimators of θ_4 and θ_5 are

$$\hat{\theta}_4 = \frac{\sum_{j\in A}\{\delta_j(y_{j1}-\bar{y}_1)(y_{j2}-\bar{y}^*_{2I}) + (1-\delta_j)\sum_{i\in A_g}\delta_i w^*_{ij}(y_{j1}-\bar{y}_1)(y_{i2}-\bar{y}^*_{2I})\}}{\sum_{j\in A}(y_{j1}-\bar{y}_1)^2},$$

and

$$\hat{\theta}_5 = \frac{\sum_{j\in A}\{\delta_j(y_{j1}-\bar{y}_1)(y_{j2}-\bar{y}^*_{2I}) + (1-\delta_j)\sum_{i\in A_g}\delta_i w^*_{ij}(y_{j1}-\bar{y}_1)(y_{i2}-\bar{y}^*_{2I})\}}{\{\sum_{i\in A}(y_{i1}-\bar{y}_1)^2\}^{1/2}[\sum_{j\in A}\{\delta_j(y_{j2}-\bar{y}^*_{2I})^2 + (1-\delta_j)\sum_{i\in A_g}\delta_i w^*_{ij}(y_{i2}-\bar{y}^*_{2I})^2\}]^{1/2}},$$

where $\bar{y}^*_{2I}$ is a mean of imputed samples in y_2. In addition to point estimators, we also computed variance estimators using the replication method in Section 3. For variance estimation, we used (4.30) for the FI estimator.

Table 4.3 presents the Monte Carlo means, variances of the point estimators and relative biases for the variance estimators. All point estimators are nearly unbiased. Slight biases in estimation of regression slope and correlation are due to discrete approximation. This bias decreases if imputation cell size increases. Ratios, $V(\hat{\theta}_{FEFI})/V(\hat{\theta}_{Full})$ and $V(\hat{\theta}_{FI})/V(\hat{\theta}_{Full})$ for (θ_1, θ_2) under $n = 300$ are $(0.77, 0.73)$ and $(0.75, 0.73)$. This implies that we have some efficiency gains considering missing rate in estimation of θ_1 and θ_2. For domain estimation, FEFI and FI estimator are more efficient than the Full sample estimator because it borrows strength outside domain (Kim and Fuller, 2004). These efficiency gains of FEFI estimators depend on correlation of two variables. Efficiency losses of FI estimator with respect to FEFI estimator are relatively small but it is almost 4% for regression slope. These efficiency losses decreases if the imputation cell size m increases. The relative biases of variance estimators are negligible ($\leq 5\%$).

In the second simulation, two-stage cluster sampling is considered. We first generated a finite population that consists of 200 clusters. The population size N is $\sum_{i=1}^{200} C_i = 39,856$, where C_i is a size of i-th cluster that was randomly generated from Poission distribution with a parameter $\lambda = 200$. From the population, we select c simple random samples of clusters and then randomly selected $n_j = 10$ samples without replacement within each sampled cluster. Sampling weights, $(200/c)(C_j/n_j)$ $(j = 1, \ldots, c)$, are assigned to final samples. Variables (Y_1, Y_2, δ_2, D) are generated with the same way of the first simulation in each cluster. We

Table 4.3 Monte Carlo results for the first simulation

Parameter	Estimator	n=300			n=500		
		Mean	Variance	R.B(%)	Mean	Variance	R.B(%)
	FULL	2.00	0.00267		2.00	0.00162	
θ_1	FEFI	2.00	0.00346	2.4	2.00	0.00208	1.9
$E(Y_2)$	FI(m=4)	2.00	0.00354	2.5	2.00	0.00213	2.1
	FULL	0.50	0.00079		0.50	0.00049	
θ_2	FEFI	0.50	0.00108	2.4	0.50	0.00064	2.9
$P(Y_2 < 2)$	FI(m=4)	0.50	0.00108	2.8	0.50	0.00064	3.3
	FULL	2.00	0.00909		2.00	0.00549	
θ_3	FEFI	2.00	0.00868	3.3	2.00	0.00534	0.1
$E(Y_2 \mid D = 1)$	FI(m=4)	2.00	0.00905	2.3	2.00	0.00553	-0.1
	FULL	1.00	0.00482		1.00	0.00315	
θ_4	FEFI	0.99	0.00710	2.0	0.99	0.00432	-1.5
Slope	FI(m=4)	0.99	0.00742	-2.4	0.99	0.00450	-4.3
	FULL	0.63	0.00101		0.63	0.00065	
θ_5	FEFI	0.62	0.00146	4.3	0.62	0.00091	-1.4
Correlation	FI(m=4)	0.62	0.00149	1.4	0.62	0.00092	-3.1

applied categorization on Y_1 and Y_2 so that Y_1 has five groups and Y_2 has two groups. From categorization of Y_2, we have $m = 4$ imputed values for each nonresponding unit. We consider two cases $c = 30, 50$ and then we have two sample sizes, $n = 300$ and $n = 500$. Also, we generated $B = 2000$ Monte Carlo samples. Variance estimation is also implemented using the replication estimator (4.28) for FI within each cluster.

Table 4.4 shows the Monte Carlo means, variances of the point estimators and relative biases for the variance estimators in the two-stage cluster sampling. All estimators are nearly unbiased. There are slight efficiency gain for variance of fractional imputation estimator with respect to the missing rate. Ratios, $V(\hat{\theta}_{FEFI})/V(\hat{\theta}_{Full})$ and $V(\hat{\theta}_{FI})/V(\hat{\theta}_{Full})$, for mean and proportion under $n = 300$ are $(0.79, 0.74)$ and $(0.77, 0.74)$. These efficiency gains are also observed for $n = 500$. For domain mean, regression slope and correlation, we have the similar results with the first simulation. Variances of FI estimator are almost as efficient as variance of FEFI estimator within 3%. Also, relative biases of variance estimators are all less than 5%.

Table 4.4 Monte Carlo results for the second simulation

Parameter	Estimator	n=300 (c=30)			n=500 (c=50)		
		Mean	Variance	R.B(%)	Mean	Variance	R.B(%)
	Full	2.00	0.00276		2.00	0.00177	
θ_1	FEFI	2.00	0.00346	2.4	2.00	0.00220	-3.9
$E(Y_2)$	FI (m=4)	2.00	0.00357	1.5	2.00	0.00226	-4.2
	Full	0.50	0.00080		0.50	0.00052	
θ_2	FEFI	0.50	0.00108	2.5	0.50	0.00069	-4.1
$P(Y_2 < 2)$	FI (m=4)	0.50	0.00108	2.4	0.50	0.00069	-4.0
	Full	2.00	0.00959		2.00	0.00561	
θ_3	FEFI	2.00	0.00911	-0.8	2.00	0.00530	1.7
$E(Y_2 \mid D = 1)$	FI (m=4)	2.00	0.00938	-0.6	2.00	0.00547	1.7
	Full	1.00	0.00511		1.00	0.00299	
θ_4	FEFI	0.99	0.00704	4.8	0.99	0.00425	0.2
Slope	FI (m=4)	0.99	0.00726	1.6	0.99	0.00434	-2.2
	Full	0.63	0.00106		0.63	0.00061	
θ_5	FEFI	0.63	0.00149	2.2	0.63	0.00086	3.7
Correlation	FI (m=4)	0.63	0.00152	0.4	0.63	0.00088	1.0

4.5.2 Multivariate case

Now we extend the proposed method to a multivariate missing case. We generated $Y_i = (Y_{1i}, Y_{2i})$, $i = 1, \cdots, n$, from

$$\begin{aligned} Y_1 &\sim U(0, 2), \\ Y_2 &= 1 + Y_1 + e_2, \\ Y_3 &= 2 + Y_1 + 0.5Y_2 + e_3 \end{aligned}$$

where e_2 and e_3 are independently generated from a normal distribution, $N(0, 1/2)$ for e_2 and $N(0, 1)$ for e_3. We generated $\delta_{ik} \sim Bernoulli(p_k)$ independently for each Y_k with $p_1 = 0.5$, $p_2 = 0.7$ and $p_3 = 0.9$ so that all variables are subject missingness.

In this simulation, the categorization process is applied to guarantee that the number of donor in each imputation cell is at least two. Each variable is firstly categorized into three groups and then collapsed into two groups corresponding the cell size requirements introduced in Appendix E. The number of imputed values depends on the missing pattern of recipient.

If all variables are respectively categorized three cells and we have fully observed units for any combination of cells, then there are 18 $(= 3 \times 3 \times 2)$ imputed values for a recipient that has singly observed item among three variables. We consider two sample sizes $n = 300$ and $n = 500$ and $B = 2,000$ Monte Carlo samples are generated.

We computed estimators of $\theta_1 = E(Y_1), \theta_2 = E(Y_2), \theta_3 = E(Y_3), \theta_4 = P(Y_1 < 1, Y_2 < 2)$ and $\theta_5 = E(Y_2 \mid D = 1)$ with $D \sim Bernoulli(0.3)$. Similar to the previous simulations, three estimators (Full, FEFI, FI with $m_h = 2$) are considered. For variance estimation, we also used a replication estimator, (4.21), for the FEFI method and used two replication estimators, (4.21) and (4.28), for the FI method.

Table 4.5 Monte Carlo results for three estimators in multivariate case.

Parameter	Estimator	n=300			n=500		
		Mean	Var.	R.B(%)	Mean	Var.	R.B(%)
	Full	1.00	0.00112		1.00	0.00067	
θ_1	FEFI	1.00	0.00188	-2.9	1.00	0.00109	-1.7
$E(Y_1)$	FI	1.00	0.00192	-2.8	1.00	0.00112	-1.7
	Full	2.00	0.00199		2.00	0.00116	
θ_2	FEFI	2.00	0.00256	0.8	2.00	0.00153	0.4
$E(Y_2)$	FI	2.00	0.00261	0.8	2.00	0.00156	0.5
	Full	4.00	0.00604		4.00	0.00358	
θ_3	FEFI	4.00	0.00657	0.3	4.00	0.00381	3.6
$E(Y_3)$	FI	4.00	0.00662	0.2	4.00	0.00383	3.4
	Full	0.40	0.00080		0.40	0.00050	
θ_4	FEFI	0.40	0.00119	5.1	0.40	0.00077	-3.2
$P(Y_1 < 1, Y_2 < 2)$	FI	0.40	0.00119	5.7	0.40	0.00077	-3.4
	Full	4.00	0.02098		4.00	0.01235	
θ_5	FEFI	4.00	0.02018	-1.0	4.00	0.01169	1.8
$E(Y_2 \mid D = 1)$	FI	4.00	0.02044	-1.6	4.00	0.01176	1.7

Table 4.5 presents the Monte Carlo means, variances of the point estimators and relative biases of the variance estimator for multivariate case. All estimators are nearly unbiased and the proposed FEFI and FI estimator perform well in this simulation. In estimation of mean estimators in θ_1, θ_2 and θ_3, ratios of variance of full sample estimators to variance of fractional imputation estimators (0.60, 0.92, 0.78) with $n = 300$ are greater than each missing rate p_k

$(k = 1, 2, 3)$. These efficiency gains are owing to correlation of variables. All FI estimators are almost as efficient as and the relative biases of variance estimators are less than 5.7% for $n = 300$ and negligible $(\leq 3.6\%)$ for $n = 500$.

4.6 Concluding remarks

In this paper, we develop a fractional hot deck imputation that does not require the imputation cell information known in advance. Basically, imputation procedure mimics two-phase stratified sampling in the sense that all imputation cells, under the cell mean model and MAR condition, are fractionally assigned to missing items of a recipient and then imputed values are jointly generated within imputed cell.

The proposed method is extended to the multivariate missing case with arbitrary missing patterns. For multivariate imputation, we need joint cell probabilities that are used to get conditional cell probabilities corresponding fraction of imputation cells. The joint distribution of the study vector is approximated by a discrete approximation. The choice for the optimal level of discrete approximation can be viewed as a bandwidth selection in nonparametric procedure. A modified EM algorithm is introduced for computation of joint cell probabilities.

One desirable feature of the proposed method is that the covariance structure of multivariate variables are retained after imputation because imputed values are jointly generated and are selected to mimic distribution of variables as closely possible by using an efficient sampling algorithm such as systematic PPS sampling. While the proposed FI estimator is nearly as efficient as the FEFI estimator, the size of the finally imputed data set will be relatively small compared to the use of FEFI. This feature will be another merit in real field. An **R** software package of the proposed method is under development.

Appendix

A. Proof of Theorem 1

Before we prove Theorem 1, assume that δ_i $(i = 1, \ldots, N)$ is extended to the entire population and assumed to be independent random variable. This extension has been discussed by Fay (1991) and used by Rao and Shao (1992).

Now, applying Taylor expansion on the $\hat{Y}_{FEFI}$ defined in (4.14), we have

$$\begin{aligned}
\hat{Y}_{FEFI} &= \sum_{g=1}^{G} \hat{R}_g^{-1} \sum_{i \in A_g} w_i \delta_i y_i = \sum_{g=1}^{G} \hat{N}_g (\hat{Y}_{Rg} / \hat{N}_{Rg}) \\
&= \sum_{g=1}^{G} \frac{N_g}{N_{Rg}} Y_{Rg} + \sum_{g=1}^{G} \frac{N_g}{N_{Rg}} (\hat{Y}_{Rg} - Y_{Rg}) \\
&\quad + \sum_{g=1}^{G} \frac{Y_{Rg}}{N_{Rg}} (\hat{N}_g - N_g) - \sum_{g=1}^{G} \frac{Y_{Rg} N_g}{N_{Rg}^2} (\hat{N}_{Rg} - N_{Rg}) + S_n + G_n, \qquad \text{(A.1)}
\end{aligned}$$

where

$$\begin{aligned}
S_n &= \frac{1}{N_{Rg}} (\hat{Y}_{Rg} - Y_{Rg})(\hat{N}_g - N_g) - \frac{N_g}{N_{Rg}^2} (\hat{Y}_{Rg} - Y_{Rg})(\hat{N}_{Rg} - N_{Rg}) \\
&- \frac{Y_{Rg}}{N_{Rg}^2} (\hat{N}_{Rg} - N_{Rg})(\hat{N}_g - N_g) + \frac{Y_{Rg} N_g}{N_{Rg}^3} (\hat{N}_{Rg} - N_{Rg})^2,
\end{aligned}$$

and G_n is a remainder term.

From the assumption $(A4)$, S_n has the order of $O_p(n^{-1}N)$. Thus, by the assumption $(A4)$ and $(A5)$, (A.1) can be expressed with

$$\hat{Y}_{FEFI} = \sum_{g=1}^{G} \sum_{i \in A_g} w_i \{ y_i + (R_g^{-1} \delta_i - 1)(y_i - \mu_g) \} + o_p(n^{-1/2} N), \qquad \text{(A.2)}$$

where, $R_g = N_{Rg}/N_g$ and $\mu_g = E(y_i) = E(\sum_{h=1}^{H} z_{ih} y_i) = \sum_{h=1}^{H} \pi_{h|g} \mu_{gh}$ for $i \in U_g$. Henceforth, we define $\tilde{Y}_{FEFI} = \sum_{g=1}^{G} \sum_{i \in A_g} w_i \gamma_{ig}$ with $\gamma_{ig} = y_i + (R_g^{-1} \delta_i - 1)(y_i - \mu_g)$.

Thus, from (4.15) and $\tilde{Y}_{FEFI}$, we have the result (4.16),

$$\hat{Y}_{FI} = \tilde{Y}_{FI} + o_p(n^{-1/2} N), \qquad \text{(A.3)}$$

where $\tilde{Y}_{FI} = \tilde{Y}_{FEFI} + \sum_{g=1}^{G} \sum_{i \in A_g} w_i (1 - \delta_i) \sum_{h=1}^{H} \hat{\pi}_{h|g} (\bar{y}^*_{i|gh} - \bar{y}_{Rgh})$.

Let $E_I(\cdot)$ be an expectation on imputation mechanism, we have

$$E_I(\tilde{Y}_{FI}) = \tilde{Y}_{FEFI}. \tag{A.4}$$

Thus, to prove (4.17), it suffices to show that $E(\tilde{Y}_{FEFI} - Y_N) = 0$.

Taking expectation on $\tilde{Y}_{FEFI}$, we have

$$\begin{aligned} E(\tilde{Y}_{FEFI}) &= E\left\{E(\bar{Y}_{FEFI} \mid \mathcal{F}_N)\right\} \\ &= E\left(\sum_{g=1}^{G}\sum_{i\in U_g} y_i\right) + E\left(\sum_{g=1}^{G}\sum_{i\in U_g}(R_g^{-1}\delta_i - 1)(y_i - \mu_g)\right) \\ &= E(Y_N) + E\left(\sum_{g=1}^{G}\sum_{i\in U_g}(R_g^{-1}\delta_i - 1)(y_i - \mu_g)\right), \end{aligned} \tag{A.5}$$

where $\mathcal{F}_N$ is a set of finite population.

On the other hand,

$$\begin{aligned} & E\left(\sum_{g=1}^{G}\sum_{i\in U_g}(R_g^{-1}\delta_i - 1)(y_i - \mu_g)\right) \\ = \; & E\left\{\sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{i\in U_g}(R_g^{-1}\delta_i - 1)z_{ih}E(y_i - \mu_g \mid x_i = g, z_{ih} = 1, \delta_i = 1)\right\} \\ = \; & E\left\{\sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{i\in U_g}(R_g^{-1}\delta_i - 1)E(z_{ih} = 1 \mid x_i = g, \delta_i = 1)(\mu_{gh} - \mu_g)\right\} \\ = \; & E\left\{\sum_{g=1}^{G}\sum_{i\in U_g}(R_g^{-1}\delta_i - 1)\sum_{h=1}^{H}\pi_{h|g}(\mu_{gh} - \mu_g)\right\} = 0, \end{aligned} \tag{A.6}$$

where second equality comes from the cell mean model and MAR assumptions in $(A2)$ and last equality comes from $\mu_g = \sum_{h=1}^{H}\pi_{h|g}\mu_{gh}$ and $\sum_{h=1}^{H}\pi_{h|g} = 1$.

From (A.4), (A.5) and (A.6), we have $E(\tilde{Y}_{FI} - Y_N) = 0$, that is, (4.17) is established.

Also, from expression (A.2), we have

$$\begin{aligned} V\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\gamma_{ig}\right) &= V\left[\sum_{g=1}^{G}\sum_{i\in A_g}\{w_i\mu_g + R_g^{-1}\delta_i(y_i-\mu_g)\}\right] \\ &= V\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g\right) + V\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i R_g^{-1}\delta_i(y_i-\mu_g)\right\} \\ &+ \text{Cov}\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g, \sum_{g=1}^{G}\sum_{i\in A_g} w_i R_g^{-1}\delta_i(y_i-\mu_g)\right\}. \end{aligned} \tag{A.7}$$

For the second term of (A.7), we have

$$\begin{aligned} & V\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i R_g^{-1}\delta_i(y_i-\mu_g)\right\} \\ = \; & E\left[\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i R_g^{-1}\delta_i(y_i-\mu_g)\right\}^2\right] - \left[E\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i R_g^{-1}\delta_i(y_i-\mu_g)\right\}\right]^2 \\ = \; & E\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i^2 R_g^{-2}\delta_i(y_i-\mu_g)^2\right\}, \end{aligned} \tag{A.8}$$

where second equality comes from $E\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i R_g^{-1}\delta_i(y_i-\mu_g)\right\} = 0$.

For the third term of (A.7), we write

$$\begin{aligned} & \text{Cov}\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g, \sum_{g=1}^{G}\sum_{i\in A_g} w_i R_g^{-1}\delta_i(y_i-\mu_g)\right\} \\ & = E\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g \sum_{g=1}^{G}\sum_{i\in A_g} w_i R_g^{-1}\delta_i(y_i-\mu_g)\right\} \\ & = E\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i^2\mu_g R_g^{-1}\delta_i(y_i-\mu_g)\right\} \\ & = E\left\{\sum_{g=1}^{G}\sum_{i\in U_g} w_i\mu_g R_g^{-1}\delta_i(y_i-\mu_g)\right\} = 0 \end{aligned} \tag{A.9}$$

where second equality comes from $E\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i R_g^{-1}\delta_i(y_i-\mu_g)\right\} = 0$, third equality comes from independence condition in (4.1) and last equality holds due to the cell mean model in (4.1).

From (A.7)-(A.9), we write

$$V(\tilde{Y}_{FEFI}) = V\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g\right) + E\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i^2 R_g^{-2}\delta_i(y_i-\mu_g)^2\right\} \tag{A.10}$$

Note that the variance of $N^{-2}\tilde{Y}_{FEFI}$ converges to the variance of $N^{-2}\hat{Y}_{FEFI}$ as n goes to infinity such that

$$\begin{aligned}
& V\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\gamma_{ig}\right) \\
= \; & E\left[\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i\gamma_{ig} - E\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\gamma_{ig}\right)\right\}^2\right] \\
= \; & E\left[\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i\{\hat{R}_g^{-1}\delta_i(y_i-\mu_g) + \delta_i(R_g^{-1}-\hat{R}_g^{-1})(y_i-\mu_g)\}\right\}^2\right] \\
= \; & E\left[\left\{\sum_{g=1}^{G}\hat{R}_g^{-1}\sum_{i\in A_g} w_i\delta_i(y_i-\mu_g)\right\}^2\right] \\
& + E\left[\left\{\sum_{g=1}^{G}(R_g^{-1}-\hat{R}_g^{-1})\sum_{i\in A_g} w_i\delta_i(y_i-\mu_g)\right\}^2\right] + (\text{Cross-product term}) \qquad \text{(A.11)} \\
= \; & V\left(\sum_{g=1}^{G}\hat{R}_g^{-1}\sum_{i\in A_g} w_i\delta_i y_i\right) + O(n^{-3/2}N^2) \\
= \; & V\left(\hat{Y}_{FEFI}\right) + o(n^{-1}N^2). \qquad \text{(A.12)}
\end{aligned}$$

The second term of (A.11) converges to 0 with order of $O(n^{-2}N^2)$ and the cross-product term converges to 0 with order of $O(n^{-3/2}N^2)$ by the condition $(A4)$ and the Schwarz inequality.

Thus, from (A.10) and (A.12), we show (4.19) such that

$$\begin{aligned}
V(\hat{Y}_{FEFI}) &= V\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\gamma_{ig}\right) + o_p(n^{-1}N^2) \\
&= V\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g\right) + E\left\{\sum_{g=1}^{G}\sum_{i\in A_g} w_i^2 R_g^{-2}\delta_i(y_i-\mu_g)^2\right\} + o_p(n^{-1}N^2).
\end{aligned} \tag{A.13}$$

We now write,

$$V\left(\hat{Y}_{FI}-\hat{Y}_{FEFI}\right) = V\left\{E_I\left(\hat{Y}_{FI}-\hat{Y}_{FEFI}\right)\right\} + E\left\{V_I\left(\hat{Y}_{FI}-\hat{Y}_{FEFI}\right)\right\},$$

where $\hat{Y}_{FI} - \hat{Y}_{FEFI} = \sum_{g=1}^{G} \sum_{i \in A_g} w_i(1-\delta_i) \sum_{h=1}^{H} \hat{\pi}_{h|g}(\bar{y}^*_{i|gh} - \bar{y}_{Rgh})$ and $V_I(\cdot)$ is a variance on imputation mechanism.

Since $\bar{y}^*_{i|g} = \sum_{h=1}^{H} \hat{\pi}_{h|g} \bar{y}^*_{i|gh}$ can be viewed as stratified sampling mean and $E_I(\bar{y}^*_{i|gh}) = \bar{y}_{Rgh}$, we have

$$V_I(\hat{Y}_{FI} - \hat{Y}_{FEFI}) = \sum_{g=1}^{G} \sum_{i \in A_g} w_i^2(1-\delta_i) \sum_{h=1}^{H} \hat{\pi}^2_{h|g}(M_{gh}^{-1} - n_{Rgh}^{-1})\hat{S}^2_{gh}, \tag{A.14}$$

where $\hat{S}^2_{gh}$ is given by $\hat{v}_{gh}^{-1} \sum_{i \in A_g} w_i \delta_i z_{ih}(y_i - \bar{y}_{Rgh})^2$ with $\hat{v}_{gh} = \sum_{i \in A_g} w_i \delta_i z_{ih}$, and

$$E_I(\hat{Y}_{FI} - \hat{Y}_{FEFI}) = 0. \tag{A.15}$$

Note that imputed values are treated as simple random samples because $y^*_{i|gh}$ is drawn with proportional to w_i in each cell.

Thus, by the result of (A.14) and (A.15),

$$V\left(\hat{Y}_{FI} - \hat{Y}_{FEFI}\right) = E\left\{ \sum_{g=1}^{G} \sum_{i \in A_g} w_i^2(1-\delta_i) \sum_{h=1}^{H} \hat{\pi}^2_{h|g}(M_{gh}^{-1} - n_{Rgh}^{-1})\hat{S}^2_{gh} \right\}. \tag{A.16}$$

Also, since $E_I(\hat{Y}_{FI}) = \hat{Y}_{FEFI}$, we have

$$\begin{aligned} Cov(\hat{Y}_{FI} - \hat{Y}_{FEFI}, \hat{Y}_{FEFI}) &= E\left\{\hat{Y}_{FEFI}\left(\hat{Y}_{FI} - \hat{Y}_{FEFI}\right)\right\} - E\left(\hat{Y}_{FI} - \hat{Y}_{FEFI}\right) E\left(\hat{Y}_{FEFI}\right) \\ &= E\left[E_I\left\{\hat{Y}_{FEFI}\left(\hat{Y}_{FI} - \hat{Y}_{FEFI}\right)\right\}\right] \\ &- E\left\{E_I\left(\hat{Y}_{FI} - \hat{Y}_{FEFI}\right)\right\} E\left(\hat{Y}_{FEFI}\right) \\ &= E\left(\hat{Y}^2_{FEFI} - \hat{Y}^2_{FEFI}\right) = 0. \end{aligned} \tag{A.17}$$

Therefore, by (A.16) and (A.17), (4.18) is established

B. Proof of Theorem 2

To show the asymptotic consistency of the replication variance estimator $\hat{Y}_{FI,1}$, we first write again $\hat{Y}_{FEFI}$ with the result (A.2),

$$\hat{Y}_{FEFI} = \tilde{Y}_{FEFI} + o_p(n^{-1/2}N).$$

We now consider the replication estimator in (4.21) with the expression (4.15),

$$\begin{aligned}\hat{Y}_{FI,1}^{(k)} - \hat{Y}_{FI} &= \sum_{g=1}^{G} \hat{R}_g^{-1(k)} \sum_{i \in A_g} w_i^{(k)} \delta_i y_i - \sum_{g=1}^{G} \hat{R}_g^{-1} \sum_{i \in A_g} w_i \delta_i y_i \\ &= \sum_{g=1}^{G} \hat{N}_g^{(k)} (\hat{Y}_{Rg}^{(k)} / \hat{N}_{Rg}^{(k)}) - \sum_{g=1}^{G} \hat{N}_g (\hat{Y}_{Rg} / \hat{N}_{Rg}). \end{aligned} \tag{B.1}$$

Applying Taylor expansion again on (B.1), by the conditions $(A4)$, $(A5)$ and $(A7)$, we have

$$\begin{aligned}&\sqrt{c_k}(\hat{Y}_{FI,1}^{(k)} - \hat{Y}_{FI}) \\ =\ & \sqrt{c_k} \sum_{g=1}^{G} \left\{ \frac{\hat{N}_g}{\hat{N}_{Rg}} (\hat{Y}_{Rg}^{(k)} - \hat{Y}_{Rg}) + \frac{\hat{Y}_{Rg}}{\hat{N}_{Rg}} (\hat{N}_g^{(k)} - \hat{N}_g) - \frac{\hat{Y}_{Rg} \hat{N}_g}{\hat{N}_{Rg}^2} (\hat{N}_{Rg}^{(k)} - \hat{N}_{Rg}) \right\} + o_p(n^{-1}N) \\ =\ & \sqrt{c_k} \sum_{g=1}^{G} \sum_{i \in A_g} (w_i^{(k)} - w_i) \gamma_{ig} + o_p(n^{-1}N) \end{aligned} \tag{B.2}$$

where the main term of (B.2) is $O_p(n^{-1}N)$. Thus, we have

$$\sum_{k=1}^{L} c_k \left(\hat{Y}_{FI,1}^{(k)} - \hat{Y}_{FI} \right)^2 = \sum_{k=1}^{L} c_k \left\{ \sum_{g=1}^{G} \sum_{i \in A_g} (w_i^{(k)} - w_i) \gamma_{ig} \right\}^2 + o_p(n^{-1}N^2). \tag{B.3}$$

Because γ_{ig} satisfies $(A8)$ with $\tau \geq 2$, by the assumption $(A6)$, the replicate estimator of $\bar{Y}_{FEFI}$ satisfies

$$\hat{V}(\tilde{Y}_{FEFI}) = V(\tilde{Y}_{FEFI} \mid \delta, \mathcal{F}_N) + o_p(n^{-1}N^2), \tag{B.4}$$

where

$$\hat{V}(\bar{Y}_{FEFI}) = \sum_{k=1}^{L} c_k \left\{ \sum_{g=1}^{G} \sum_{i \in A_g} (w_i^{(k)} - w_i) \gamma_{ig} \right\}^2.$$

We write the variance of $\tilde{Y}_{FEFI}$ can be expressed as

$$V(\tilde{Y}_{FEFI} \mid \mathcal{F}_N) = V\{E(\tilde{Y}_{FEFI} \mid \delta, \mathcal{F}_N) \mid \mathcal{F}_N\} + E\{V(\tilde{Y}_{FEFI} \mid \delta, \mathcal{F}_N) \mid \mathcal{F}_N\},$$

and

$$\begin{aligned} V(\tilde{Y}_{FEFI}) &= E\{V(\tilde{Y}_{FEFI} \mid \mathcal{F}_N)\} + V\{E(\tilde{Y}_{FEFI} \mid \mathcal{F}_N)\} \\ &= E[V\{E(\tilde{Y}_{FEFI} \mid \delta, \mathcal{F}_N) \mid \mathcal{F}_N\}] + E\{V(\tilde{Y}_{FEFI} \mid \delta, \mathcal{F}_N)\}. \end{aligned} \tag{B.5}$$

Note that $V\{E(\tilde{Y}_{FEFI} \mid \mathcal{F}_N)\} = 0$. From the result (B.4), $\hat{V}(\tilde{Y}_{FEFI})$ is approximately unbiased for the second term of the right-hand side of (B.5).

On the other hand, we have

$$E\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i p_g^{-1}\delta_i(y_i-\mu_g) \mid \delta, \mathcal{F}_N\right) = \sum_{g=1}^{G}\sum_{i=1}^{N} p_g^{-1}\delta_i(y_i-\mu_g)$$

and

$$V\left(\sum_{g=1}^{G}\sum_{i=1}^{N} p_g^{-1}\delta_i(y_i-\mu_g) \mid \mathcal{F}_N\right) = \sum_{g=1}^{G}\sum_{i=1}^{N} V(p_g^{-1}\delta_i)(y_i-\mu_g)^2,$$

by independence of (δ_i, x_i, y_i). Then, we have

$$E\left\{V\left(\sum_{g=1}^{G}\sum_{i=1}^{N} p_g^{-1}\delta_i(y_i-\mu_g) \mid \mathcal{F}_N\right)\right\} = \sum_{g=1}^{G}\sum_{i=1}^{N} V(p_g^{-1}\delta_i)\sigma_g^2, \tag{B.6}$$

where $\sigma_g^2 = \sum_{h=1}^{H}\pi_{g|h}(\sigma_{gh}^2+\mu_{gh}^2) - \sum_{h=1}^{H}\sum_{s=1}^{H}\pi_{h|g}\pi_{s|g}\mu_{gh}\mu_{gs}$.

Therefore, combining the results (B.3)-(B.6), we have

$$\hat{V}(\hat{Y}_{FE,1}) = V(\hat{Y}_{FEFI}) - \sum_{g=1}^{G}\sum_{i=1}^{N} V(p_g^{-1}\delta_i)\sigma_g^2 + o_p(n^{-1}N^2), \tag{B.7}$$

where

$$\hat{V}(\hat{Y}_{FE,1}) = \sum_{k=1}^{L} c_k\left(\hat{Y}_{FI,1}^{(k)} - \hat{Y}_{FI}\right)^2.$$

Now we consider the second term of the replication estimator in (4.31). For simplicity, we assume $M_{gh}=2$. Because $\phi_{gh}^{(q,s)}=0$ for $g\neq q$ or $h\neq s$, we write

$$\begin{aligned}\hat{Y}_{FI,2}^{(q,s)} - \hat{Y}_{FI} &= \sum_{g=1}^{G}\sum_{i\in A_g} w_i(1-\delta_i)\sum_{h=1}^{H}\hat{\pi}_{h|g}(\bar{y}_{i|gh}^{*(q,s)} - \bar{y}_{i|gh}^{*}) \\ &= \sum_{i\in A_q} w_i(1-\delta_i)\hat{\pi}_{s|q}\phi_{qs}(y_{i|qs}^{*(1)} - y_{i|qs}^{*(2)})\zeta_i^{(q,s)}.\end{aligned}$$

Since $\zeta_i^{(q,s)}$ is independently and equally distributed for -1 and 1,

$$E_I(\hat{Y}_{FI,2}^{(q,s)} - \hat{Y}_{FI}) = 0, \tag{B.8}$$

and

$$\begin{aligned}V_I(\hat{Y}_{FI,2}^{(q,s)} - \hat{Y}_{FI}) &= \sum_{i\in A_q} w_i^2(1-\delta_i)\hat{\pi}_{s|q}^2 M_{qs}\hat{f}_{Rqs}\phi_{qs}^2\hat{S}_{qs}^2 \\ &= \sum_{i\in A_q} w_i^2(1-\delta_i)\hat{\pi}_{s|q}^2(M_{qs}^{-1} - n_{Rqs}^{-1})\hat{S}_{qs}^2.\end{aligned} \tag{B.9}$$

where ϕ_{qs} is a solution of $M_{qs}\hat{f}_{Rqs}\phi_{qs}^2 = (M_{qs}^{-1} - n_{Rqs}^{-1})$, Thus, by the result of (B.10) and (B.9), we have

$$\begin{aligned} & E\left\{\sum_{q=1}^{G}\sum_{s=1}^{H}(\hat{Y}_{FI,2}^{(q,s)} - \hat{Y}_{FI})^2\right\} \\ = \; & V\left\{\sum_{q=1}^{G}\sum_{s=1}^{H} E\left(\hat{Y}_{FI,2}^{(q,s)} - \hat{Y}_{FI} \mid \delta\right)\right\} + E\left\{\sum_{q=1}^{G}\sum_{s=1}^{H} V\left(\hat{Y}_{FI,2}^{(q,s)} - \hat{Y}_{FI} \mid \delta\right)\right\} \\ = \; & E\left\{\sum_{q=1}^{G}\sum_{i\in A_q} w_i^2(1-\delta_i)\sum_{s=1}^{H}\hat{\pi}_{s|q}^2(M_{qs}^{-1} - n_{Rqs}^{-1})\sigma_{qs}^2\right\}. \end{aligned} \tag{B.10}$$

Therefore, by the results (B.7) and (B.10), the result (4.31) is established.

C. Description of the EM algorithm

The EM algorithm is used here in a slightly modified way. For each unit i, the conditional probability of $\tilde{\mathbf{y}}_{i,mis}$ given $\tilde{\mathbf{y}}_{i,obs}$ is computed using the current estimate of the joint probability $\tilde{\pi}(\tilde{\mathbf{y}})$, where $\sum_{\tilde{\mathbf{y}}}\pi(\tilde{\mathbf{y}}) = 1$. This is the E-step of the EM algorithm. The initial conditional probability is computed by

$$w_{ij}^* = \frac{p_0(\tilde{\mathbf{y}}_{i,obs}, \tilde{\mathbf{y}}_{i,mis} = \tilde{\mathbf{y}}_{mis(i),j})}{\sum_{k=1}^{H_i} p_0(\tilde{\mathbf{y}}_{i,obs}, \tilde{\mathbf{y}}_{mis(i),k})}. \tag{C.1}$$

where $p_0(\tilde{\mathbf{y}})$ is the estimated joint probability computed from the full respondents and $\tilde{\mathbf{y}}_{mis(i),k}$ is the k-th possible vector for the missing part of unit i. Here, without loss of generality, we assume H_i support vectors for $\tilde{\mathbf{y}}_{i,mis}$. The M-step computes the joint probability of particular combination of $\tilde{\mathbf{y}} = \tilde{\mathbf{y}}^*$ by

$$\pi(\tilde{\mathbf{y}}^*) = (\sum_{i=1}^{n} w_i)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{H_i} w_i w_{ij}^* I(\tilde{\mathbf{y}}_{i,obs} = \tilde{\mathbf{y}}_{obs(i)}^*, \tilde{\mathbf{y}}_{mis(i),j} = \tilde{\mathbf{y}}_{mis(i)}^*), \tag{C.2}$$

where $(\tilde{\mathbf{y}}_{obs(i)}^*, \tilde{\mathbf{y}}_{mis(i)}^*)$ are partitions of $\tilde{\mathbf{y}}^*$ based on the observed part and the missing pattern of unit i. Thus, equations (C.1) and (C.2) form a set of iterative computations for the EM algorithm. In the modified EM, we first compute w_{ij}^* and then update π.

Now, we can use $p^{(t)}(\tilde{\mathbf{y}}^*)$ in (C.2) to denote the t-th iteration of the computation for the joint probability so that (C.2) becomes

$$\pi^{(t+1)}(\tilde{\mathbf{y}}^*) = (\sum_{i=1}^{n} w_i)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{H_i} w_i w_{ij}^{*(t)} I(\tilde{\mathbf{y}}_{i,obs} = \tilde{\mathbf{y}}_{obs(i)}^*, \tilde{\mathbf{y}}_{mis(i),j} = \tilde{\mathbf{y}}_{mis(i)}^*),$$

where

$$w_{ij}^{*(t)} = \frac{p^{(t)}(\tilde{\mathbf{y}}_{i,obs}, \tilde{\mathbf{y}}_{i,mis} = \tilde{\mathbf{y}}_{mis(i),j})}{\sum_{k=1}^{H_i} p^{(t)}(\tilde{\mathbf{y}}_{i,obs}, \tilde{\mathbf{y}}_{i,mis} = \tilde{\mathbf{y}}_{mis(i),k})}.$$

In the classical EM algorithm, the choice of initial fractional weights are $w_{ij(0)}^{*} = 1/H_i$. In the modified EM, the initial fractional weights are the empirical conditional distribution calculated from the full respondents.

D. Description of the first replication fractional weights

To hold the expression (4.22) and the basic properties of fractional weights that they are non-negative and the sum of fractional weights over all donors is one for each recipient, the replication fractional weights $w_{ij}^{*(k)}$ can be obtained by minimizing

$$\sum_{j \in A_g} \sum_{i \in A_g} (w_{ij}^{*(k)} - w_{ij}^{*})^2 \tag{D.1}$$

subject to restrictions

$$w_{ij}^{*(k)} \geq 0, \tag{D.2}$$

$$\sum_{i \in A_g} w_{ij}^{*(k)} = 1, \tag{D.3}$$

$$\sum_{j \in A_g} w_j^{(k)} (1 - \delta_j) \sum_{i \in A_g} \delta_i (w_{ij}^{*(k)} - w_{ij}^{*}) y_i = \hat{Y}_{Mg}^{*} - \hat{Y}_{Mg}^{*(k)} + \hat{N}_{Mg}^{(k)} \bar{y}_{Rg}^{(k)} - \hat{N}_{Mg} \bar{y}_{Rg}. \tag{D.4}$$

While the restrictions (D.3) and (D.4) are satisfied by the replication fractional weights (4.26) obtained by the regression weighting method, the restriction (D.2) may not be guaranteed in a certain situation. In that case, we may apply a quadratic programming or the restriction (D.4) can be relaxed to the collapsed cell $G(g)$ including the cell g. That is, the restriction (D.4) becomes

$$\sum_{j \in A_{G(g)}} w_j^{(k)} (1 - \delta_j) \sum_{i \in_{G(g)}} \delta_i (w_{ij}^{*(k)} - w_{ij}^{*}) y_i = \hat{Y}_{MG(g)}^{*} - \hat{Y}_{MG(g)}^{*(k)} + \hat{N}_{MG(g)}^{(k)} \bar{y}_{RG(g)}^{(k)} - \hat{N}_{MG(g)} \bar{y}_{RG(g)}.$$

To illustrate the computation of the proposed replication fractional weights, we suppose two variables x and y with a size $n = 10$. Variable x is generated from a uniform distribution, $U(0, 2)$, and the study variable y given x is generated from a normal distribution $N(1 + x, 1)$. Also, y is subject missingness with a probability $p_i = 0.7$. Here, we assume single cell for x and two imputation cells

Table 4.6 An illustrative data set

id	cell.x	cell.y	y	weights
1	1	2	1.80	1
2	1	2	1.85	1
3	1	M	M	1
4	1	1	1.11	1
5	1	1	0.06	1
6	1	M	M	1
7	1	1	1.56	1
8	1	2	4.20	1
9	1	M	M	1
10	1	2	2.82	1

for y. Thus, we have $2 \times M(=2) = 4$ imputed values for each missing y. Table 4.6 shows the sample observations, where nonresponse is denoted by M in the table.

Table 4.7 shows the imputed values for each missing y with fractional weights w_{ij}^{*}. The sum of four fractional weights equals to the weight of recipient.

Table 4.8 represents the replication fractional weights obtained by using a quadratic programming method, where $w_{ij}^{**(k)} = w_{j}^{(k)} w_{ij}^{*(k)}$. In the i-th replication weight $(i = 1, 2, 3, 4)$ of the j-th recipients $(j = 1, 2, 3)$, $w_{ij}^{*(k)}$, we have $w_{ij}^{*(k)} \geq 0$, $\sum_{i=1}^{4} w_{ij}^{*(k)} = 1$ and $\sum_{j \in A_g} w_{j}^{(k)} (1 - \delta_j) \sum_{i \in A_g} \delta_i (w_{ij}^{*(k)} - w_{ij}^{*}) y_i = \hat{Y}_{Mg}^{*} - \hat{Y}_{Mg}^{*(k)} + \hat{N}_{Mg}^{(k)} \bar{y}_{Rg}^{(k)} - \hat{N}_{Mg} \bar{y}_{Rg}$ have values of $(0.09, 0.05, -0.67, 0.47, 1.05, 0.75, 0.22, -1.25, -0.24, -0.$ for $k = 1, \ldots, 10$.

Table 4.7 Fractional weights

id	cell.x	cell.y	donor.id	y	weights	w_{ij}^*
1	1	2		1.80	1	1
2	1	2		1.85	1	1
3-1	1	1	4	1.11	1	0.21
3-2	1	1	5	0.06	1	0.21
3-3	1	2	1	1.80	1	0.29
3-4	1	2	2	1.85	1	0.29
4	1	1		1.11	1	1
5	1	1		0.06	1	1
6-1	1	1	4	1.11	1	0.21
6-2	1	1	7	1.56	1	0.21
6-3	1	2	8	4.20	1	0.29
6-4	1	2	10	2.82	1	0.29
7	1	1		1.56	1	1
8	1	2		4.20	1	1
9-1	1	1	5	0.06	1	0.21
9-2	1	1	7	1.56	1	0.21
9-3	1	2	2	1.85	1	0.29
9-4	1	2	10	2.82	1	0.29
10	1	2		2.82	1	1

Table 4.8 Replication fractional weights using a quadratic programming

id	$w_{ij}^{**(1)}$	$w_{ij}^{**(2)}$	$w_{ij}^{**(3)}$	$w_{ij}^{**(4)}$	$w_{ij}^{**(5)}$	$w_{ij}^{**(6)}$	$w_{ij}^{**(7)}$	$w_{ij}^{**(8)}$	$w_{ij}^{**(9)}$	$w_{ij}^{**(10)}$
1	0	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11
2	1.11	0	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11
3-1	0.24	0.24	0	0.23	0.23	0.23	0.24	0.25	0.24	0.24
3-2	0.23	0.23	0	0.19	0.14	0.09	0.22	0.36	0.27	0.28
3-3	0.32	0.32	0	0.34	0.37	0.39	0.33	0.25	0.30	0.29
3-4	0.32	0.32	0	0.34	0.38	0.40	0.33	0.25	0.30	0.29
4	1.11	1.11	1.11	0	1.11	1.11	1.11	1.11	1.11	1.11
5	1.11	1.11	1.11	1.11	0	1.11	1.11	1.11	1.11	1.11
6-1	0.23	0.23	0.33	0.19	0.12	0	0.21	0.38	0.28	0.29
6-2	0.23	0.23	0.30	0.20	0.16	0	0.22	0.33	0.26	0.27
6-3	0.33	0.33	0.20	0.39	0.48	0	0.35	0.13	0.26	0.24
6-4	0.32	0.32	0.29	0.33	0.35	0	0.32	0.28	0.31	0.30
7	1.11	1.11	1.11	1.11	1.11	1.11	0	1.11	1.11	1.11
8	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0	1.11	1.11
9-1	0.23	0.23	0.34	0.18	0.10	0.05	0.21	0.40	0	0.30
9-2	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0	0.24
9-3	0.32	0.32	0.30	0.33	0.34	0.35	0.32	0.29	0	0.31
9-4	0.33	0.32	0.23	0.37	0.43	0.47	0.34	0.19	0	0.27
10	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0

E. Categorization algorithm

During the discretization, it is often to have zero marginal probability for $\tilde{\mathbf{y}}$ and exactly one possible donor in a cell. The zero marginal probability means that has zero probability and this implies that there is no donor in A_R. Also, we require at least $M_h(\geq 2)$ imputed values for each realized imputation cell. To avoid those problems, we can consider several discretization methods. In this paper, we introduce the following categorization procedure:

(1) Apply G_k categorization transformation on y_k, $k = 1, \ldots, p$, using the quantiles of each y_k.

(2) Obtain a frequency table for $\tilde{\mathbf{y}}_{obs}$.

(3) If the frequency table includes a cell that the number of elements equals to 1 or there exists a missing pattern that have zero marginal probability, then we apply $(G_k - 1)$ categorization transformation on the kth item and go back to (2). The kth item variable is selected by the order of missing rate and the size of categorization.

  (3.1) When the size of categorization is the same across all items, we select the variable which has the highest missing rate.

  (3.2) When the size of categorization is different from each other, then we select the variable which hasthe largest categorization size. If there are several candidates, then the variable is selected in order of missing rate.

(4) Repeat (2)-(3) until there is no case which has zero marginal probability and no cell which has only one element.

F. Program design

We now describe the details for creating computer programs for multivariate fractional hot deck imputation. The procedure will first be made in R and then programmed into SAS. The *input* variables for the procedure are give by

- ID
- Study Variable: $\mathrm{VAR}_1, \cdots, \mathrm{VAR}_p$

- Missing indicator variable: $\mathrm{RESP}_1, \cdots, \mathrm{RESP}_p$
- Cell variable: $\mathrm{CELL}_1, \cdots, \mathrm{CELL}_p$.
- Sampling weights: WGT
- Imputation size: $\mathrm{M}_c = 2$

The ID variable takes integer values from 1 to n. The missing indicator variables are dichotomous (taking 0 or 1 values), with $\mathrm{RESP}_j = 1$ if VAR_j is observed. The cell variables are all categorical. We assume that the cell variables are created from the nature of the study variable by the person who will perform the imputation. Note that, if VAR_j is missing then CELL_j is also missing.

The *output* variables for the procedure are give by

- ID
- Fraction ID: takes values from $\{1, \cdots, M_i\}$.
- Imputed Study Variable: $\mathrm{IVAR}_1, \cdots, \mathrm{IVAR}_p$
- Imputed Cell variable: $\mathrm{ICELL}_1, \cdots, \mathrm{ICELL}_p$.
- Fractional weights: FWGT
- Replication fractional weights: $\mathrm{RFWGT}_1, \ldots, \mathrm{RFWGT}_L$.

[Note: The imputation cells for a missing unit are determined by missing pattern of the missing unit. Thus, the total size of imputed values can be different for each missing unit.]

In the imputation procedure, the program has five parts:

[Part 1] Computing the joint cell probability using the modified EM algorithm described in Appendix C.

[Part 2] Computing the conditional cell probability:

For each unit i, compute the cell probability corresponding to the particular cell $c = (c_1, \cdots, c_p)$ by

$$\begin{aligned} \hat{\pi}_{ic} &= \frac{P(\text{CELL} = c)}{\sum_{\{k;\text{OBS}_i(c)=\text{OBS}_i(k)\}} P(\text{CELL} = k)} \text{ if } \text{OBS}_i(c) = \text{OBS}_i(\text{CELL}(i)) \\ &= 0 \text{ otherwise} \end{aligned}$$

where $\text{CELL}(i)$ is the value of $\text{CELL} = (\text{CELL}_1, \cdots, \text{CELL}_p)$ for unit i, and $\text{OBS}_i(c)$ is a function that gives the value of c for the observed part of unit i. Thus, condition $\text{OBS}_i(c) = \text{OBS}_i(\text{CELL}(i))$ means that the particular cell $c = (c_1, \cdots, c_p)$ has the same values of the observed part of CELL values for unit i.

[Note: It is possible to have $\sum_c \hat{\pi}_{ic} = 0$ for some i. In this case, we output some warning that this case does not give proper fractional weights for the given cell designation. Some cells must be collapsed and rerun the procedure to get $\sum_c \hat{\pi}_{ic} = 1$.]

[Part 3] Imputation of donors. For each imputation cell, the donors are taken by a systematic PPS sampling. The systematic PPS sampling for fractional imputation can be performed as follows.

1. Choose $R_0 \sim \text{Unif}(0, M_c^{-1} n_m^{-1})$, where n_m is the size of nonresponding units.
2. For each nonresponding unit j ($j = 1, \ldots, n_m$), we define $R_j = R_0 + M_c^{-1} n_m^{-1}$. The unit k ($k = 1, 2 \cdots, n$) is selected as donor for the nonresponding unit j if

$$\sum_{i=1}^{k-1} w_i^* < R_j + \frac{l}{M_c} \leq \sum_{i=1}^{k} w_i^*,$$

for some $l \in \{0, \cdots, M_c - 1\}$ and $w_i^* = w_i \delta_i z_{ic} / \sum_i w_i \delta_i z_{ic}$, where w_i and $\delta_i = \prod_{k=1}^{p} \delta_{ik}$ are realized value for $\text{WGT}_i$and RESP_i. Also, z_{ic} is a cell indicator function that takes the value one if unit i belongs to cell c.

[Part 4] Fractional weight. Let w_{ij}^* be estimated fractional weight for FWGT_{ij}. The fractional weight for imputation cell c is computed by

$$w_{ij}^* = \hat{\pi}_{ic} M_c^{-1}$$

where $\hat{\pi}_{ic}$ is an estimated conditional cell probability computed in [Part 2].

[Part 5] Replication weight. Let M_c be even number and $\mathbf{y} = (y_1^*, \ldots, y_n^*)$ be a final imputed data for IVAR such that

$$y_j^* = \begin{cases} y_j & \text{if} \quad \delta_i = 1 \\ (y_{j,1}^*, \ldots, y_{j,M_j}^*) & \text{otherwise,} \end{cases}$$

where $y_{j,i}^*$ is the i-th imputed value for the nonresponding unit j and $\tilde{w}^{(k)} = (w_1^{(k)}, \ldots, w_n^{(k)}$ is k-th replication fractional weights for RFWGT_k. Then, k-th $(k = 1, \ldots, L)$ replication fractional weights are given by

$$\tilde{w}_{1j}^{(k)} = \begin{cases} w_j^{(k)} & \text{if} \quad \delta_i = 1 \\ w_j^{(k)}(w_{1j,i}^{(k)}, \ldots, w_{1j,M_j}^{(k)}) & \text{otherwise,} \end{cases}$$

where $w_{1j,i}^{*(k)}$ $(i = 1, \ldots, M_j)$ are obtained using w_{ij}^* of [Part 4] and (4.15).

[Note: Since it is possible to have negative replication fractional weights, optional computation process using a quadratic programming would be incorporated in the program.]

The h-th replicates for unit corresponding to the second replicatoin estimator, we have

$$\tilde{w}_{2j}^{(h)} = \begin{cases} w_j & \text{if} \quad \delta_i = 1 \\ w_j(w_{2j,1}^{(h)}, \ldots, w_{2j,M_j}^{(h)}) & \text{otherwise,} \end{cases}$$

where $w_{2j,i}^{(h)} = w_{ij}^* + w_{ij}^{*(h)}$ with $w_{ij}^{*(h)} = \hat{\pi}_{ih}\phi_c^{(h)}\zeta_i^{(h)}q_{ij,c}$,

$$\phi_c^{(h)} = \begin{cases} \phi_c & \text{if } c = h, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$q_{ij,c} = \begin{cases} 1 & \text{for the first donor group in cell } c, \\ -1 & \text{for the second donor group in cell } c. \end{cases}$$

Also, $\zeta_i^{(h)}$ is an independent variable taking 1 or -1 with equal probability and ϕ_c is obtained by solving $M_c\hat{f}_c\phi_c^2 = M_c^{-1} - n_{Rc}^{-1}$, where $\hat{f}_c = n_{Rc}/(n_{Rc} - 1)$ and n_{Rc} is a size of responding units in cell c.

The program provides the imputed data with imputed cells and fractional weights as basic outputs and two types of replication weights for FI estimator are also provided if the FI estimator is identified as the input options.

BIBLIOGRAPHY

Aitkin, M. A. (1964). Correlation in a singly truncated bivariate normal distribution. *Psychometrika*, *29*, 263–270.

Alho, J. M. (1990). Adjusting for nonresponse bias using logistic regression. *Biometrika*, *77*, 617–24.

Amemiya, T. (1973). Regression Analysis when the dependent variable is truncated normal. *Econometrica*, *41*, 997–1016.

Arismendi, J. C. (2013). Multivariate truncated moments. *Journal of Multivariate Analysis*, *117*, 41–75.

Arnold, B. C., Beaver, R. J. and Meeker, R. A. G. (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika*, *58*, 471–488.

Azzalini, A. (1985). A class of distribution which includes normal ones. *Scandinavian Journal of Statistics*, *12*, 171–178.

Barr, D. R. and Davidson, T. G. (1973). A Kolmogorov-Smirnov test for censored examples. *Technometrics*, *15*, 739–757.

Breslow, N. E. and Cain, K. C. (1988). Logistic regression for the two-stage case-control data. *Biometrika*, *75*, 11–20.

Breslow, N. E. and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters for two-phase outcome dependent sampling. *Journal of Royal Statistical Society: Series* B,

Brick, J. M. and Kalton, G. (1996). Handling missing data in servey research. *Statistical Methods in Medical Research*, *5*, 215–238.

Cartinhour, J. (1990). One-dimensional marginal density functions of a truncated multivariate normal density function. *Communications in Statistics: Part A*, *19*, 197–203.

Chang, T. and Kott, P. (2008). Using calibration weighting to adjust for nonignorable under a plausible model. *Biometrika*, *95*, 555–71.

Chen, S. X., Leung, D. H. Y. and Qin, J. (2008). Improving semiparametric estimation by using surrogate data. *Journal of Royal Statistical Society: Series* B, *70*, 803–23.

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, *16*, 113–132.

Cohen, A. C. (1955). Restriction and selection in samples from bivariate normal distributions. *Journal of the American Statistical Association*, *50*, 884–893.

Cotton, C. (1991). *Functional description of the Generalized Edit and Imputation System*, Business Survey Methods Division, Statistics Canada.
59, 447–461.

Deming, W. E. (1953). On a probability mechanism to attain an economic balance between the resultant error of response and the bias of nonresponse. *Journal of the American Statistical Association*, *48*, 743–72.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*, 376–382.

Drew, J. H. and Fuller, W. A. (1980). Modeling nonresponse in surveys with callbacks. In *Proceeding of the Survey Research Methods Section*, American Statistical Association, pp. 639–42.

Drew, J. H. and Fuller, W. A. (1981). Nonresponse in complex multiphase surveys. In *Proceeding of the Survey Research Methods Section*, American Statistical Association, pp. 623–28.

Dufour, R. and Maag, J. R. (1978). Distribution results for modified Kolmogorov-Smirnov statistics for truncated or censored samples. *Technometrics*, *20*, 29–32.

Durbin, J. (1975). Kolmogorov-Smirnov test when parameters are estimated with application to tests of exponentiality and tests on spacings. *Biometrika*, *62*, 5–22.

Fay, R. E. (1991). A design-based perspective on missing data variance. In *Proceedings of Bureau of the Census Annual Research Conference*, American Statistical Association, 429–440.

Fleming, T. R., O'Fallon, J. R., O'Brien, P.C. and Harrington, D.P. (1980). Modified Kolmogorov-Smirnov test procedure with application to arbitrarily right-censored data. *Biometrics*, *36*, 607–625.

Folsom, R. E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. In *Proceeding of the Social Statistics Section*, American Statistical Association, pp. 197–202.

Fuller, W. A. (2009). *Sampling Statistics.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Fuller, W. A., Loughin, M. M. and Baker, H. D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology*, *20*, 75–85.

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

Goldberger, A. S. (1981). Linear regression after selection. *Journal of Econometrics*, *21*, 195–212.

Green, W. H. (1983). Estimation of limited dependent variable models by ordinary least squares and the method of moments. *Journal of Econometrics*, *21*, 195–212.

Hansen, M. H. and Hurwitz, W. N. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, *41*, 517–29.

Hazziza, D. (2009). Imputation and inference in the presence of missing data. In*Handbook of Statistics*, Volume *29*, *Sample Surveys: Theory Methods and Inference*, Edited by C.R. Rao and D. Pfeffermann, 215–246.

Hazziza, D. and Beaumont, J. F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, *75*, 25–43.

Iannacchione, V. G., Milne, J. G. and Folsom, R. E. (1991). Response probability weight adjustments using logistics regression. In *Proceeding of the Survey Research Methods Section*, American Statistical Association., pp. 637–42.

Judkins, D., Krenzke, T., Piesse, A., Fan, Z., and Haung, W. C. (2007). Preservation of skip patterns and covariate structure through semi-parametric whole questionnaire imputation. In *Proceeding of the Survey Research Methods Section*, American Statistical Association, 3211-3218.

Kalbfleisch, J. D. and Lawless, J. F. (1989). Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS. *Journal of the American Statistical Association*, *84*, 360–372.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, *12*, 1–16.

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

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

Kim,J. K., Navarro, A., and Fuller, W. A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of American Statistical Association*, *101*, 312–320.

Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, *35*, 501–14.

Kim, J. K. and Riddles, M. K. (2012). Some theory for propensity-score-adjustment estimators in survey sampling. *Survey Methodology*, *38*, 157–65.

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

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, *32*, 133–42.

Kott, P. S. and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, *105*, 1265–75.

Koziol,J. R. and Byar, D. P. (1975). Percentage points of the asymptotic distribution of one and two sample K-S statistics for truncated or censored data. *Technometric*, *17*, 507–510.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, *33*, 101–16.

Olsen, R. J. (1980). Approximating a truncated normal regression with the method of moments. *Econometrica*, *48*, 1099–1105.

Orme, C. (1989). On the uniqueness of the maximum likelihood estimator in truncated regression models. *Econometric Reviews*, *8*, 217–222.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya*, *61*, 166–186.

Pfeffermann, D. and Sikov, A. (2011). Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information. *Journal of Official Statistics*, *27*, 181–209.

Proctor, C. (1977). Two direct approaches to survey nonresponse: estimating a proportion with callbacks and allocating effort to raise the response rate. In *Proceeding of the Social Statistics Section*, American Statistical Association, pp. 284–90.

Rancourt, E., Särndal, C. E., and Lee, H. (1994). Estimation of the variance in presence of nearest neighbor imputation. In *Proceeding of the Survey Research Methods Section*, American Statistical Association, 888–893.

Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, *60*, 125–133.

Rao, B. R., Garg, M. L. and Li, C. C.(1968). Correlation between the sample variances in a singly truncated bivariate normal distribution. *Biometrika*, *55*, 433–436.

Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, *79*, 811–822.

Rizzo, L., Kalton, G. and Brick, J. M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, *22*, 44–53.

Rosenbaum, S. (1961). Moments of a truncated bivariate normal distribution. *Journal of the Royal Statistical Society: Series* B, *23*, 405–408.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–90.

Rubin, D.B. (1987). Multiple imputation for nonresponse in surveys, New York: John Wiley & Sons, Inc.

Rubin, D.B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, *81*, 366–374.

Shao, J. and Wang, H. (2002). Sample correlation coefficients based on survey data under regression imputation. *Journal of American Statistical Association*, *97*, 544–552.

Tang, G., Little, R. J. A. and Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, *90*, 747–64.

Wang, S., Shao, J. and Kim, J. K. (2014). An instrument variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, *24*, 1097–1116.

Wild, C. J. (1991). Fitting prospective regression models to case-control data. *Biometrika*, *78*, 705–717.

Wood, A M., White, I. R. and Hotopf, M. (2006). Using number of failed contact attempts to adjust for non-ignorable non-response. *Journal of Royal Statistical Society: Series* A , *169*, 525–42.