2009

Fractional Imputation

Minhui Paik
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd
Part of the Statistics and Probability Commons

Recommended Citation
Paik, Minhui, "Fractional Imputation" (2009). Graduate Theses and Dissertations. 10710.
http://lib.dr.iastate.edu/etd/10710

This Dissertation is brought to you for free and open access by the Graduate College 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.
Fractional Imputation

by

Minhui Paik

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:
Michael D. Larsen, Major Professor
Wayne A. Fuller
Frederick Lorenz
Jae-Kwang Kim
Daniel Russell

Iowa State University
Ames, Iowa
2009

Copyright © Minhui Paik, 2009. All rights reserved.
DEDICATION

I would like to dedicate this thesis to God. Without his guidance and grace, this thesis would not be done. Also, I am grateful to dedicate this thesis to my family, my wife, Sujin, my daughter, Annette and my parents. I would not have been able to complete this work without their support and encouragement.

I would like to thank my supervisor, Dr. Michael Larsen, for his mentoring through my PhD studies. He showed me his great patience and encouragement when my progress was at a slow pace, which is not of time. This thesis would not be finished if he had not been advising and helping. Also his passion for research and teaching will always lead me to dream my future career. I would like to thank him for the financial support during my studies, too.

I also would like to thank my committee member, Dr. Jae-Kwan Kim for many good suggestions to improve this thesis. He would inspire me to gain better understanding of the subject areas in this these and help me to go through some of the hardest parts of the proof and developments. In addition I want to thank Dr. Wayne Fuller for the suggestion of some important points in an earlier version of this thesis. Lastly, I thank my committee member, Dr. Fred Lorenz, to allow me to use the data sets for enlarging the usefulness of this research.
TABLE OF CONTENTS

LIST OF TABLES ......................................................... vi

ABSTRACT ................................................................. viii

INTRODUCTION ............................................................ 1

FRACTIONAL DETERMINISTIC IMPUTATION ......................... 4
  Abstract ........................................................................ 4
  1. Introduction ........................................................... 5
  2. Imputation Methods .................................................. 11
     2.1. Hot Deck Imputation ............................................ 12
     2.2. Nearest Neighbor Imputation ................................. 15
     2.3. Regression Imputation ........................................... 17
     2.4. Fractional Imputation ............................................ 20
     2.5. Multiple Imputation: MI ........................................ 30
     2.6. Comparison of Fractional Hot Deck Imputation (FHDI) and MI .... 34
     2.7. Fractional Regression Hot Deck Imputation: FRHDI ........... 35
     2.8. Fractional Nearest Neighbor Imputation: FNNI ................ 39
  3. Some theoretical results for imputation methods ............... 40
     3.1. Basic Setup ...................................................... 40
     3.2. Order in probability .............................................. 42
     3.3. Results for the Linear Regression Imputation Estimator ........... 45
     3.4. Results Concerning Fractional Imputation Estimators .......... 47
WEIGHT ADJUSTMENTS FOR FRACTIONAL REGRESSION HOT DECK IMPUTATION

Abstract .......................................................................................... 123
1. Introduction .................................................................................. 123
2. Fractional Regression Hot Deck Imputation ................................. 126
3. Construction of Regression Fractional Weights ......................... 129
   3.1. Introduction to Regression Weighting ................................. 129
   3.2. Computer Algorithm for Regression Fractional Weights ....... 131
4. Simulation Study .......................................................................... 134
5. Summary and Discussion ............................................................. 138

BIBLIOGRAPHY .............................................................................. 144
## LIST OF TABLES

| Table A.1 | Monte Carlo means and variances based on 5,000 replicates for point estimators of the mean, the domain mean, and a proportion using various methods under the first simulation conditions | 91 |
| Table A.2 | Monte Carlo means and variances based on 5,000 replicates for point estimators of correlation and variance using various methods under the first simulation conditions | 92 |
| Table A.3 | Relative biases, t-statistics, and relative variance for variance estimators of the mean, the domain mean, and a proportion using various methods under the first simulation conditions | 93 |
| Table A.4 | Relative biases, t-statistics, and relative variance for variance estimators of correlation and variance using various methods under the first simulation conditions | 94 |
| Table A.5 | Monte Carlo means and variances based on 5,000 replicates for point estimators of the mean, correlation, and variance using various methods under the second simulation conditions | 95 |
| Table A.6 | Relative biases, t-statistics, and relative variance for variance estimators of the mean, correlation and variance using various methods under the second simulation conditions | 96 |
| Table B.1 | Point estimators for Trivariate Normal Sample with Missing Data on Two Variables | 121 |
Table B.2  Variance estimators for Trivariate Normal Sample with Missing Data on Two Variables .......................... 122
Table C.1  Monte Carlo mean and variance of the point estimator under simulation 1 .......................... 140
Table C.2  Monte Carlo relative biases, t-statistics and relative variances of the variance estimator under simulation 1 .......................... 141
Table C.3  Monte Carlo mean and variance of the point estimator under simulation 2 .......................... 142
Table C.4  Monte Carlo relative biases, t-statistics and relative variances of the variance estimator under simulation 2 .......................... 143
ABSTRACT

Sample surveys typically gather information on a sample of units from a finite population and assign survey weights to the sampled units. Surveys frequently have missing values for some variables for some units. Imputation is widely used in sample surveys as a method of handling the missing data problem.

We provide a new imputation procedure using empirical likelihood to provide an easy-to-use data set for general purpose and keep the desirable properties of deterministic imputation method under fractional imputation. The imputed estimator constructed by the proposed procedure is called the Fractional Deterministic Imputation (FDI) estimator. The construction of the FDI method is discussed in detail in order to describe the general proposed procedure. In addition, a computationally efficient variance estimator is given that permits the construction of general purpose replicates for variance estimation. In order to deal with multivariate missing data, the proposed imputation method can be extended via calibration to match results obtained using maximum likelihood estimation for some parameters under multivariate normal distribution model.

Finally, we address an issue common to several weight adjustment methods; namely, the issue of highly variable or even negative weights. Adapting an existing algorithm, our modification provides solutions with positive weights within a bounded interval.

The combined contributions of this dissertation extend methods of imputation for missing values in important ways. Methods can be used with both complex survey and other data.
INTRODUCTION

Sample surveys typically gather information on a sample of units from a finite population and assign survey weights to the sampled units. Surveys frequently have missing values for some variables for some units. Imputation is widely used in sample surveys as a method of handling the missing data problem. Imputation methods generally require assumptions of two types in order to be appropriate for use with a particular study variable. One assumption is expressed by the imputation model and the other is expressed by the response model. The imputation model makes an assumption about the distribution of the study variable and its relationship with other variables collected in the survey. The other model is called the response model and concerns assumptions about the probability of obtaining respondents from the sample. Analyses incorporating the imputation model and the response model are quite popular in practice even though in cases of unintended missing data it usually is hard to defend assumed models as being more realistic than other candidate models.

An ideal objective for an imputation procedure is that it be applicable in general settings and that estimators involving the imputation procedure have good properties. Ideally the estimators perform well for analyses that are planned before imputation and also for analyses that are unplanned at the time the imputation model is created. The main objective of this dissertation is to describe and illustrate a fractional imputation method that yields an easy-to-use data set and that comes closer to the ideal objective than currently existing methods.

Specifically, in a first paper, we develop a new imputation method that should be
of practical use in general situations and for which both theoretical and practical requirements are satisfied. The proposed method is an imputation procedure for various imputation models retaining many of the desirable properties of model-based imputation and hot-deck imputation within the framework of fractional imputation. For the completed data set constructed by the proposed procedure, the standard estimates at the aggregate level of analysis are equivalent to model-based imputation estimates for parameters in the imputation model. Other population parameters not included in the imputation procedure can also be estimated efficiently. A computationally efficient variance estimator is developed to provide the construction of general purpose replicates for variance estimation. The imputed estimator constructed by the proposed procedure is called the Fractional Deterministic Imputation (FDI) estimator. The construction of the FDI estimator is discussed in detail in order to describe the general proposed procedure.

Following the first paper, a section containing theoretical results for FDI is presented. This section concerns the consistency of the FDI estimator and its replication variance estimator.

In a second paper, fractional imputation ideas are extended via calibration to maximum likelihood estimates to situations with general patterns of missing data. This development allows fractional imputation methods using various methods for creating imputations or choosing donors to be used when data are missing on more than one variable. The illustration provided uses multivariate normal data, but the idea can be applied potentially to other situations.

In a third paper, an iterative regression procedure is investigated for the purpose of constructing non-negative factional weights and to place bounds on the fractional weights. This weight adjustment procedure can be used with several versions of fractional imputation. It is important because some fractional imputation methods can produce non-positive weights, which are not practically useful for variance estimation. They also can yield extreme weights, which lead to extreme standard errors and large estimated
standard errors.

A conclusion section to the dissertation provides a summary of the contributions of the doctoral research and comments on directions for future development.
FRACTIONAL DETERMINISTIC IMPUTATION

A paper will appear as a proceedings paper of the 2009 Joint Statistical Meetings.

Minhui Paik and Michael D. Larsen

Abstract

Sample surveys typically gather information on a sample of units from a finite population and assign survey weights to the sampled units. Surveys frequently have missing values for some variables for some units. Imputation is a widely used method in sample surveys as a method of handling the missing data problem.

We provide an new imputation procedure for various imputation models retaining many of the desirable properties of deterministic imputation under fractional imputation. The main objective of this procedure is to construct an easy-to-use data set for general purpose analysis. The imputed estimator constructed by the proposed procedure is called the Fractional Deterministic Imputation (FDI) estimator. The construction of the FDI estimator is discussed in detail in order to describe the general proposed procedure. In addition, a computationally efficient variance estimator is given to provide the construction of general purpose replicates for variance estimation.

A simulation study is conducted to compare the performance of FDI with other imputation methods including fractional nearest neighbor imputation (FNNI, Fuller and Kim, 2005a), fractional regression hot deck imputation (FRHDI, Kim, 2007) and multiple imputation (MI, Rubin, 1978).
1. Introduction

Nonresponse in one form or another is frequently present in most surveys. Nonresponse means failure to obtain a measurement on one or more study variables for one or more elements of the population selected for the survey. Nonresponse is a serious issue in survey sampling and in statistics in general. As stated in the Statistics Canada Quality Guidelines (Statistics Canada 2003):

*Non-response leads to an increase in variance as a result of a reduction in the actual size of the sample and the recourse to imputation. This produces a bias if the non-respondents have characteristics of interest that are different from those of the respondents. Furthermore, there is a risk of significantly underestimating the sampling error, if imputed data are treated as though they were observed data.*

There are two types of non-response. First, a sampled unit that is contacted may fail to respond. This represents “unit nonresponse”. However, there is often some information available such as the general location or strata information on the nonresponding units. Second, the unit may respond to the questionnaire incompletely. This is referred to as “item nonresponse”. We call a unit with an observed value a *respondent* and call a unit with a missing value a *nonrespondent*.

Often a data set that has missing values can be viewed as being produced through the process of two-phase sampling according to which the set of respondents is treated as a second phase sample from the original sample. The two-phase design, by design, has missing information for first-phase units that are not included in the second-phase sample. When an actual two-phase sample is taken, such missing data are called planned or designed missingness. When missingness is unplanned, unlike in real two-phase sampling, and there is nonresponse, then the probability of obtaining responses from the sample is unknown and response models often can be estimated using the information contained in auxiliary variables.
A common method for addressing the issue of missing data is to increase the sampling weights of the respondents using estimated response probabilities. Response probabilities are usually estimated within strata or domains of the population. Separate and possibly differential adjustments to sampling weights are made within subgroups. Doing so should to some degree compensate for potential nonresponse bias in the sample survey. This method is called “nonresponse weighting adjustment” (NWA). The NWA estimator for a total is analogous to the Horvitz-Thompson estimator for a total but with the inverse probability of selection weights replaced by the nonresponse adjusted weights. The properties of the NWA estimators using the estimated response probabilities have been studied by several authors including Rosenbaum (1987), Robins, Rotnitzky and Zhao (1994), Little and Vartivarian, and Beaumont (2005), and Kim and Kim (2007). They noted that the NWA estimator using the estimated response probabilities can be more efficient than the NWA estimator using the true response probabilities. Specifically, Beaumont (2005) and Kim and Kim (2007) gave a clear justification for the reduction in variance using estimated response probabilities and the asymptotic properties of this estimator. Unfortunately, the NWA estimator is very sensitive to misspecification of the response model. A simple but strong assumption for a response model called the response homogeneity groups (RHG) model is commonly used. The RHG model assumes that the population consists of nonoverlapping subgroups, often called cells, and all units in the same cell have the same response probability. Even if the model were exactly true, one would need to use the correct cells in order to apply it and effectively remove all bias.

Another method of implementing estimation in data sets with missing data is imputation. Imputation is a commonly used method of compensating for item nonresponse in sample surveys, especially when the amount of missing information is rather small. Imputation methods replace missing values by feasible and plausible values. The values that replace the missing values are called imputed values. Imputing all the missing val-
ues produces a completed data set that should be as easy to analyze as a complete data set even for users who do not have special familiarity with analyzing non-rectangular data. An advantage of imputing the missing values is that one data set is produced and different users can analyze the data set. The results obtained by different users should be consistent with one another. If each user applies different adjustments for the missing data, then different users could produce different results for the same analysis when there are missing data. There are many possible imputation procedures. To design an imputation procedure, we suggest following criteria or goals. An imputation procedure should

A.1 Improve the efficiency of estimators of parameters and compensate for nonresponse bias,

A.2 Have available a valid variance estimation procedure for use with the imputed estimator, and

A.3 Be appropriate for planned estimates as well as unplanned estimates.

The key to a successful imputation procedure for nonresponse lies in the use of auxiliary information since the effective use of auxiliary information will reduce both the nonresponse bias and increase the efficiency of estimators. Detailed use of auxiliary information is demonstrated, for example, in Särndal and Lundström (2005). Thus, developing a method for incorporating information contained in auxiliary variables into the imputed estimator is an important consideration when designing an imputation procedure.

Criteria (A.2) is also important. It is well known that if the standard variance estimator is applied to an imputed complete data set, it may lead to serious underestimation of the true variance of the imputed estimator since the imputed values are treated as if they were true values. For solving this underestimation problem, various variance
estimation methodologies have been suggested. They include multiple imputation (Rubin, 1978), adjusted jackknife variance estimation (Rao and Shao, 1992) and fractional imputation with replication variance estimation (Kim and Fuller, 2004), respectively. The fractional imputation procedure and variance estimation procedure given by Kim and Fuller (2004) is useful for general purposes since in their construction of replicates, when constructing replicates, only the weights on the imputed values are changed. This means that the replicate weights can be easily used for estimators of smooth function of the response variable mean without recomputing imputed values.

Criteria (A.3) is related to the practical issue of designing an imputation procedure for actual applications. Two cases can be considered in practice. One is a situation in which the statistician knows the major estimators that will be constructed from a data set before conducting an imputation procedure. The other case is one in which the full set of possible estimators is not known. In the former situation, the advantages of an imputation procedure in terms of producing an easy-to-use data set is not as great. A model for the distribution of variables and response probability will be developed for the parameters of known interest. One could simply use the estimator to directly obtain estimates from the model. For example, if the parameter of interest is only the mean of a variable and one assumes that there is a uniform response probability, then the most efficient estimator of the mean is simply the mean of the responding units. For producing an easy-to-use data set, a corresponding imputation procedure that produces the same estimate is to replace missing values with the respondents’ mean. An alternative that produces nearly the same estimate is to randomly pick donor values for the missing values from the respondent values. However, the typical survey situation is that one makes available to different users a completed data set without knowing what kind of parameters the users of the data want to estimate. If some users want to estimate other parameters than the mean, such as quantiles or regression coefficients, the mean-imputed data set will produce estimates that will be seriously biased. Hence, a ideal objective is
to design an imputation procedure for general purpose estimation as anticipated in the latter situation.

According to our criteria for designing an imputation procedure, one can consider the fractional imputation (FI) procedure suggested by Kim and Fuller (2004). Fractional imputation procedures impute more than one value for each missing value and assign fractional survey weights to the multiply imputed values. First, the FI method was developed to reduce the imputation variance which came from the random component of the variance of the estimator arising from imputation. This procedure was suggested by Kalton and Kish (1984). Later, Kim and Fuller (2004) developed this procedure under hot deck imputation for general purpose use in survey sampling.

One alternative procedure consists of assigning every respondents to each nonrespondent and giving equal weight to each donor value. In most cases, defining cells and taking donors within the cell makes the most sense. Kim and Fuller (2004) called this procedure fully efficient fractional imputation in that there is no variance due to the imputation conditional on the sample. In the resulting completed data sets, the estimate for any function of a response variable equals the value that would be obtained by using only the values from the sample respondents. Fully efficient fractional imputation may not be used so commonly because of the size of the resulting data set. However, Fuller and Kim (2005a) also proposed very efficient procedures when using a small number of donors. In their study, fractional hot deck imputation is more efficient than multiple imputation based on the same number of donors in a particular situation. They also suggest a consistent replication variance estimation procedure for their fractional hot deck method. As we mentioned, their replication variance estimation is appropriate for general purposes because it can be easily extended to variance estimation for nonlinear statistics without changing the imputed values themselves. The sense in which this extension is valid will be described in the next section.
Of course there are limitations to any method. If the model used for imputation, which could be defined implicitly through the cell structure and rules for donating cases, is very wrong, then the fractionally imputed estimator could have problems. Additionally, hot deck imputation of missing variables on a single variable using the cell mean model may not preserve the correlation structure among two or more quantitative variables. Except for the variables that define cells, fractional hot deck imputation under the cell mean model ignores covariates. Recently, some fractional imputation procedures that incorporate the auxiliary information into the fractionally imputed estimator were developed by Fuller and Kim (2005b) and Kim (2003, 2007). These developments will be described in the next section.

In this paper, we provide an imputation procedure for various imputation models retaining many of the desirable properties of deterministic imputation method. The main objective of this procedure is to construct an easy-to-use data set that satisfies the criteria. For the completed data set constructed by the proposed procedure, the standard estimates at the aggregate level of analysis are equivalent to deterministic imputation estimates for parameters. For some other parameters not included in model, the imputation procedure leads to efficient estimation of these parameters as well. In addition, a computationally efficient variance estimator is given to provide the construction of general purpose replicates for variance estimation. The imputed estimator constructed by the proposed procedure is called the Fractional Deterministic Imputation (FDI) estimator. The construction of the FDI estimator is discussed in detail in order to describe the general proposed procedure.

This article is organized as follows. In Section 2, the properties of several imputation methods are discussed. Some theoretical results for imputation methods are described in Section 3. In section 4, the proposed FDI method is presented. A conclusion is given in Section 5.
2. Imputation Methods

Let $U = \{1, 2, \ldots, N\}$ be a set of indices of a finite population $F_N$ of $N$ elements with $N$ known. Associated with the $i^{th}$ unit of $U$ is $y_i$, a variable of interest, and $x_i$, an auxiliary variable (or vector of variables). Let $A$ denote the set of indices of the elements in a sample of size $n$ selected by the chosen sampling mechanism. For $i \in A$, assume that some $y_i$ are missing, but that every $x_i$ is observed. Responses $y_i$ are obtained from the selected sample according to the response mechanism. Let $A_r$ be the set of respondents and $A_m$ the set of nonrespondents. In this notation, $A = A_r \cup A_m$ and $A_r \cap A_m = \emptyset$.

Let the population quantity of interest be $\theta_N = \theta(y_1, y_2, \ldots, y_N)$. Let $\hat{\theta}_1n$ be a design linear estimator of $\theta_1N = \sum_{i=1}^{N} y_i$ based on the full sample,

$$\hat{\theta}_1n = \sum_{i \in A} w_i y_i$$

(2.1)

where $w_i = (Pr(i \in A))^{-1}$ is a sampling weight for unit $i$ that depends on the sampling mechanism. A second quantity of interest is $\theta_2N = \sum_{i=1}^{N} x_i y_i$, which is estimated in the full sample by the design linear estimator

$$\hat{\theta}_2n = \sum_{i \in A} w_i x_i y_i.$$  

(2.2)

We assume that $Pr(i \in A) > 0$ for all $i$, so that

$$E(\hat{\theta}_1n | F_N) = \theta_1N$$  

(2.3)

and

$$E(\hat{\theta}_2n | F_N) = \theta_2N$$  

(2.4)

where $E(\cdot | F_N)$ is the design expectation and denotes the average over all samples possible under the design for the given finite population $F$. The estimators $\hat{\theta}_1n$ and $\hat{\theta}_2n$ are the traditional Horvitz-Thompson estimators. If units in $A_r$ are the only ones observed, then
estimators of \( \theta_{1N} \) and \( \theta_{2N} \) that increase weights on the observed responses according to

\[
    w_i' = w_i \left( \sum_{i \in A} w_i \right) / \left( \sum_{i \in A} w_i \right)
\]

are

\[
    \hat{\theta}_{1r} = \sum_{i \in A_r} w_i' y_i
\]

(2.5)

and

\[
    \hat{\theta}_{2r} = \sum_{i \in A_r} w_i' x_i y_i,
\]

(2.6)

respectively. These estimators are called “Nonresponse weighting adjustment” (NWA) estimators.

Imputation methods fill-in values for the missing information. Suppose for the moment that one value is imputed for each missing value. Let \( y_i^* \) be the imputed value for \( y_i \) if it is missing and equal to \( y_i \) if it is observed. Let \( w_i^* \) be a weight for case \( i \) after the process of imputation has been implemented for all missing cases. It is possible that the weights will be unchanged, but one might alter weights for one or more units for various reasons. Based on the singly imputed data set, let the estimators of \( \theta_{1N} \) and \( \theta_{2N} \) be

\[
    \hat{\theta}^I_{1n} = \sum_{i \in A} w_i^* y_i^*
\]

(2.7)

and

\[
    \hat{\theta}^I_{2n} = \sum_{i \in A} w_i^* x_i y_i^*,
\]

(2.8)

respectively. Reasons and methods for altering \( w_i^* \) from the original \( w_i \) will be described in the upcoming sections.

2.1. Hot Deck Imputation

Hot deck imputation is an imputation procedure in which the value from an observed member of the sample is used as the imputed value for a missing item. An advantage of the method is that actually occurring values are used for imputation. The method therefore would be suitable for dealing with categorical data or quantitative data with
range restrictions. The unit providing the value is called the donor and the unit with the missing value is called the recipient. Many hot deck procedures select donor values within imputation cells based on the observed auxiliary variables, including variables used to define the sampling design. Selecting within imputation cells should increase the chance that a donated value is close to the actual value of the missing observation. The imputation cells could be defined for the whole population, such as cells derived by county of residence. Alternatively, imputation cells could be created once the sample is observed and some information is gathered on all sample units. In some sampling situations, auxiliary information might be available for all units on the sample even if the response $y_i$ is missing for some cases. Often hot deck imputation cells are considered fixed during variance estimation even if they are determined after sample selection.

In random hot deck imputation, the donors are selected at random for nonrespondents overall or at random within the same imputation cell. Typically one imputation is chosen for each missing value, but of course multiple random selections for each missing value could be made. Multiple selections could be made with or without replacement.

Random hot deck imputation preserves the distributional properties of the observed data $\{y_i, i \in A_r\}$ in the combined imputed and observed data set. Suppose that a single donor is selected for each missing value according to a hot deck procedure without imputation cells, weights are unchanged (e.g., $w_i^* = w_i$) after imputation, and selection is done with replacement. For simplicity, assume that donors are selected with probabilities proportional to their weights $w_i$: $w_i / \sum_{A_r} w_i$. Let $\bar{y}_{rw} = \sum_{A_r} w_i y_i / \sum_{A_r} w_i$ be the weighted mean of the respondent values. Conditional on the set of respondents and
their observed values, the imputed estimator on average equals the NWA estimator:

\[
E(\hat{\theta}_1|A_r) = E(\sum_{A_r} w_i y_i + \sum_{A_m} w_i y_i^* | A_r) \tag{2.1.1}
\]

\[
= \sum_{A_r} w_i y_i + \sum_{A_m} w_i E(y_i^* | A_r)
\]

\[
= \sum_{A_r} w_i y_i + \sum_{A_m} w_i \bar{y}_{rw}
\]

\[
= (\sum_{A_r} w_i + \sum_{A_m} w_i) \bar{y}_{rw},
\]

\[
= \sum_{A_r} w_i y_i,
\]

(2.1.2)

which is the same as the estimator \( \hat{\theta}_1 \) above. It can be similarly shown that \( E(\hat{\theta}_2|A_r) = \hat{\theta}_2 \).

The random selection increases the variance of an estimator relative to an estimator constructed from only the observed data set. This can be seen by viewing the variance of \( \hat{\theta}_1 \) (or \( \hat{\theta}_2 \)) as the expectation of a conditional variance plus the variance of a conditional mean. In formulas, this can be written as

\[
V(\hat{\theta}_1) = E[V(\hat{\theta}_1|A_r)] + V[E(\hat{\theta}_1|A_r)] \tag{2.1.3}
\]

\[
= E[\sum_{A_m} w_i^2 V(y_i^* | A_r)] + V(\hat{\theta}_1).
\]

The first term of the last line above is not less than zero.

Brick and Kalton (1996) describe two methods for reducing imputation variance due to random selection of donors. One method is to use a more efficient sampling method such as simple random sampling without replacement, stratified sampling, or systematic sampling to select the donors for the set of missing values within imputation cells. That is, one could restrict the imputations so that each respondent is used only once as a donor or the set of donors is intentionally spread more evenly throughout the observable cases. A second approach is to use fractional imputation which involves using more than
one donor for a recipient. More details for fractional imputation are reported in Section 2.4. The goal of the present work is to propose and evaluate an improved fractional imputation procedure.

2.2. Nearest Neighbor Imputation

Nearest neighbor imputation (NNI) can be considered a special case of the hot deck procedure in which the donor is selected by minimizing the distance to the recipient based on a suitable distance measure. The suitable distance measure can be constructed as a function of the auxiliary variables. The observed unit with a smallest distance to the nonrespondent unit is used as a donor. If multiple donors are needed, then the closest several potential donors can be selected.

Chen and Shao (2000) summarize some of the advantages of the NNI method. First, it is a hot deck method in the sense that missing items are replaced by the values for observed units so that the imputed values are actually occurring values and are not artificially constructed values. Second, the imputed values may not be perfect substitutes for the true responses, but they are unlikely to be nonsensical values. Third, since the NNI method uses the information in the auxiliary variables, the NNI method may be more efficient than other hot deck imputation schemes that either do not use auxiliary information or use it less precisely. Fourth, as with other hot deck procedures, NNI makes no distributional assumptions except implicitly through the choice of matching variables. In comparison, imputation methods such as regression imputation, which is discussed in Section 2.3, make explicit model assumptions such as the assumption of a linear regression model.

Chen and Shao (2000) prove results concerning the imputation estimator in the case of nearest neighbor imputation (NNI). In order to prove results, the authors make assumptions about the distributions of the matching variable $x$ in the respondent and nonrespondent populations. They also make assumptions about the relationship between
$y$ and $x$ in the superpopulation, or process that generated the finite population under study. If the relationship between $y$ and $x$ is linear and distributions of $x$ are symmetric for respondents and nonrespondents, then the NNI estimator is unbiased. If, as is typical, the distribution of $x$ is not symmetric, then there is bias of order $r^{-1}$, where $r$ is the size of $A_r$. Assuming the number of respondents increases with sample size, the NNI estimator is asymptotically unbiased. The authors also produce approximate variance formulas and propose a model-based variance estimator.

Chen and Shao (2001) study several jackknife variance estimators for NNI estimator. They showed that the naive jackknife estimator underestimates, whereas Rao and Shao (1992)’s adjusted jackknife estimator overestimates. In their paper, two partially reimputed jackknife estimators and one partially adjusted jackknife estimator, which is asymptotically unbiased and consistent, are proposed. However, when the sampling fraction $n/N$ is not negligible, the proposed jackknife variance estimators tend to be underestimate.

It is common practice to use a single donor using the NNI method and a distance measure defined on the basis of only observed auxiliary variables. This is one of the motivations for fractional nearest neighbor (FNNI) method (Fuller and Kim 2005b) in Section 2.8. The idea of Fuller and Kim (2005b) is to use two or more donors selected by NNI method for each missing unit so that it is easy to represent uncertainty due to imputation without special variance estimation formulas such as those in Chen and Shao (2000, 2001). Further it would be desirable to be able to accommodate a variety of survey sampling designs without deriving new variance estimation results for each design. With missing data on several variables, marginally imputing them using the NNI method may result in biased estimators for estimating a measure of relationship between two variables. In this case, bias correction should be made before variance estimation is done. Skinner and Rao (2002) proposed the method using an adjustment to eliminate biases.
2.3. Regression Imputation

Another classical method for imputing missing data is regression imputation. Consider the following model:

\[ E(y_i|x_i) = \beta_0 + \beta_1 x_i \]  \hspace{1cm} (2.3.1)

and

\[ V(y_i|x_i) = \sigma^2. \]

In regression imputation, the cases in \( A_r \) with both \( x \) and \( y \) observed are used to estimate the regression model parameters \((\beta_0, \beta_1, \sigma^2)\). In plain regression imputation, the missing values are replaced by a predicted value \( y_i^* = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \), where \( x_i \) is the observed \( x \)-value for the case with a missing outcome and \((\hat{\beta}_0, \hat{\beta}_1)\) are estimated regression coefficients. In many practical cases, the missing value is replaced by a value predicted by an estimated regression model plus a random residual, drawn to reflect uncertainty in the predicted value. In that case, the imputed value \( y_i^* \) is

\[ y_i^* = \hat{y}_i + \hat{\epsilon}_i. \]  \hspace{1cm} (2.3.2)

The residual terms are added to allow for natural randomness so that the imputations will not distort the observed conditional distribution shape of the study variables and will reflect inherent uncertainty in \( y \). Adding residuals is a common practice (see Rubin 1987, Rubin and Schenker 1986, and references below).

The residuals can be obtained in at least two alternative ways: i) by drawing residual values from a normal distribution with mean zero and estimated standard deviation \((\hat{\sigma}^2)\), or ii) by selecting randomly from the set of empirical residual \( \{\hat{\epsilon}_i = y_i - \hat{y}_i, i \in A_r\} \). Method ii) could be preferred when some assumptions of the regression models are not completely reasonable. For example, one could imagine that someone would fit a constant variance linear regression model to data for which the residual variance actually
increases or changes with $x$. Doing so can produce unbiased estimates of regression slopes, but the estimators are not as efficient as if the variance structure had been incorporated into estimation. If residuals are drawn within cells defined by $x$, then the imputed values will roughly mimic the heteroscedasticity of the observed values. Another situation would arise when the linear regression model is a reasonable approximation to the actual process that generated the data, but slight nonlinearities are present. As a result, for some intervals of $x$, the residuals could tend to be mostly positive or mostly negative. Drawing residuals from the observed empirical set of residuals within intervals defined by $x$ would tend to preserve the observed patterns in the residuals.

Deville and Särndal (1994) considered the properties of regression imputation without the addition of residuals. Under the assumption that the response mechanism is unconfounded and the imputation model is correct, the imputation estimator is unbiased under the combined regression-sampling-response model. An unconfounded or ignorable response mechanism is one for which the probabilities of response do not depend on the value of the variable that would be seen with response. The authors also propose a variance estimation alternative that incorporates the uncertainty due to imputation. It seems that the methods of this paper would need to be redone for each estimand of interest.

Särndal and Lundström (2005; pages 161-163) comment on the methods, including noting that imputation without adding a residual might be more efficient for point estimation than adding a residual, but for variance estimation and preserving the pattern in the data doing so is not advocated. Srivastava and Carter (1986) in their study of maximum likelihood estimation for survey data in the presence of nonresponse derive estimators under an assumption of multivariate normality of observations. This leads to regression imputation. The authors mention that imputing values on the regression line or plane will lead to underestimation of variability and that one can add randomly drawn empirical residuals to produce more realistic variability.
Shao and Wang (2002), extending work by Srivastava and Carter (1986), consider both plain regression imputation and random regression imputation that adds residuals drawn using method (i). The authors develop a method for joint imputation of two variables with missing values that preserves correlations. In order to estimate variances the authors extend the adjusted jackknife variance estimation method of Rao and Shao (1992). The adjustments would need to be made for each regression model of interest.

In general, variance estimation with single imputation can be a challenge, as noted by several of the authors cited above. Single imputation cannot represent uncertainty due to imputation. Variance estimation tends to underestimate variance except in special situations using special formulas, such as those of Rao and Shao (1992).

In general, plain regression imputation should preserve the regression of $y$ on $x$, because all imputed points will be on the regression line. Imputation on the regression line of course will increase the apparent correlation. Random regression imputation should preserve the regression of $y$ on $x$ and also the correlation between the two variables. In terms of estimating the mean of $y$ both regression imputation methods should provide unbiased estimates of the mean under the assumption of ignorable nonresponse and the regression model. Regression imputation should have smaller imputation variance than a simple hot deck procedure due to the use of the auxiliary information.

The performance of regression imputation and nearest neighbor imputation could be quite similar in most scenarios. Regression imputation could be superior if the model is correct, but nearest neighbor imputation could provide for more robust estimates under departures from the regression model. Despite the advantages of a well-chosen regression model and imputation scheme for imputing missing values, imputing a single value for each missing value, still has a disadvantage. Except in a few specific estimations, it is unlikely that the variance of the resulting estimator will be correctly estimated. That is, using imputed data as if they are real tends to lead to understatement of variability (Little and Rubin 2002; Deville and Särndal 1994, page 386).
A simple, naive example illustrates the situation. Assume $r$ respondents have $(x_i, y_i)$ observed, but $m$ ($r + m = n$) nonrespondents have only $x_i$ observed. Suppose the $m$ nonrespondents have the values $y_i^* = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ imputed. Define $y_i^{**} = y_i$ if $y_i$ is observed and $y_i^{**} = y_i^*$ otherwise. Then the usual least squares estimate of residual variance is too small:

$$
\hat{\sigma}_I^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i^{**} - \hat{y}_i)^2
$$

$$
= \frac{1}{n-2} \sum_{i=1}^{r} (y_i - \hat{y}_i)^2,
$$

$$
< \frac{1}{r-2} \sum_{i=1}^{r} (y_i - \hat{y}_i)^2 = \hat{\sigma}_{cc}^2.
$$

(2.3.3)

The last line would be the appropriate estimator of variability if the values are missing completely at random and only the $r$ complete cases were used. Even if the larger estimate of residual variability $\hat{\sigma}_{cc}^2$ is used, a naive approach will underestimate the variability of the slope estimator. Assuming $\bar{x}_n \approx \bar{x}_{cc}$,

$$
\hat{V}(\hat{\beta}_1) = \left( \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \right)^{-1} \hat{\sigma}^2
$$

$$
< \left( \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \right)^{-1} \hat{\sigma}_{cc}^2
$$

$$
< \left( \sum_{i=1}^{r} (x_i - \bar{x}_{cc})^2 \right)^{-1} \hat{\sigma}_{cc}^2.
$$

(2.3.4)

2.4. Fractional Imputation

Fractional imputation (FI) selects multiple donors for each missing observation and assigns a weight equal to a fraction of the original survey weight for each donor. The method of FI was originally suggested by Kalton and Kish (1984) as a method for improving the efficiency of the imputed point estimator by reducing variance (conditional on an observed sample) due to imputation. Later it was shown by Kim and Fuller (2004)
to provide a useful tool for variance estimation. Their study showed some advantages of FI over some other methods.

Let $d_{ij}$ be the number of times that unit $i$ is used as a donor for the missing unit $j$ and define $\mathbf{d} = \{d_{ij} : i \in A_r, j \in A_m\}$. The distribution of $\mathbf{d}$ is called the imputation mechanism. If selection of donors is done without replacement from the set of respondents, then values of $d_{ij}$ are zero or one. If selection is done with possible replacement, then $d_{ij}$ could be more than 1. Selection of donors without replacement generally will have less imputation variance than selection with replacement, so typically without replacement would be preferred. In some cases, however, there might be very few plausible donors for some cases, in which case selection without replacement could force the use of donors that do not match the case with missing values very well.

Let $w_{ij}^*$ be the factor applied to the original weight for missing unit $j$ when unit $i$ is used as a donor for element $j$. The factor $w_{ij}^*$ is called a fractional weight. The weight from unit $j$ ‘given’ to donor $i$ is then $w_j w_{ij}^*$. The sum of fractional weights for a missing item is restricted to equal one; i.e., for $j \in A_m$,

$$\sum_{i \in A_{Dj}} w_{ij}^* = 1,$$

where $A_{Dj}$ is the set of indices of imputed values for element $j$. The set $A_{Dj}$ is constructed by an imputation mechanism where $A_{Dj} = \{i : d_{ij} \geq 1, i \in A_r\}$. If selection with replacement is allowed, then for donors to case $j$, ($i \in A_{Dj}$), it is assured that the same weight is used each time case $i$ denotes its value to missing element $j$. Then

$$\sum_{i \in A_{Dj}} w_{ij}^* = \sum_{i \in A_r} d_{ij} w_{ij}^{**} = 1,$$

where for $d_{ij} \geq 1$, $w_{ij}^{**} = w_{ij}^*/d_{ij}$.

For element $j$, $j \in A_m$, the weighted mean of the imputed values is

$$y_{ij}^* = \sum_{i \in A_r} d_{ij} w_{ij}^* y_i.$$
Under fractional imputation, the imputed estimators of $\theta_1$ and $\theta_2$ can be written in the following forms:

$$\hat{\theta}_{1n} = \sum_{i \in A_r} w_i y_i + \sum_{j \in A_m} w_j y_{ij}$$

$$= \sum_{i \in A_r} w_i y_i + \sum_{j \in A_m} w_j \sum_{i \in A_{Dj}} w_{ij}^* y_i$$

$$= \sum_{i \in A_r} w_i y_i + \sum_{j \in A_m} \sum_{i \in A_r} w_{ij} d_{ij} w_{ij}^* y_i$$

$$= \sum_{i \in A_r} \left( \sum_{j \in A_m} \left( w_i + \sum_{j \in A_m} w_{ij} d_{ij} w_{ij}^* \right) y_i \right)$$

$$= \sum_{i \in A_r} w_i^* y_i,$$

(2.4.4)

and

$$\hat{\theta}_{2n} = \sum_{i \in A_r} \left( w_i x_i + \sum_{j \in A_m} w_j d_{ij} w_{ij}^* x_j \right) y_i$$

$$= \sum_{i \in A_r} w_i^* y_i,$$

(2.4.5)

respectively. Note that the notation $A =: B$ means that $B$ is defined to be equal to $A$.

The modified weight for respondent $i$ is $w_i^* = w_i + \sum_{j \in A_m} w_j d_{ij} w_{ij}^*$ for $\hat{\theta}_{1n}$ and $w_{x,i}^* = w_i x_i + \sum_{j \in A_m} w_j d_{ij} w_{ij}^* x_j$ for $\hat{\theta}_{2n}$.

When $A_{Dj}$ is equal to $A_r$, the estimators (2.4.4) and (2.4.5) with constraint (2.4.1) and $d_{ij} \leq 1$ are called fully efficient fractionally imputed (FEFI) estimators. They are fully efficient in the sense that, given the set $A_r$, they have no variance due to imputation because they use all observed cases once as donors for each missing unit. Note that it is not the case that the average imputed values are the same for all missing cases under fully efficient fractional imputation. The fractional weights $w_{ij}^*$ depend on both the missing case $j$ and the donor case $i$, so that using all the possible donors as donors does not produce only one average value. In particular, as will be seen, the fractional weights $w_{ij}^*$ can be related to values on auxiliary variable $x_i$ and $x_j$. 
There are various ways to implement fractional imputation. In this section, the fractional hot deck imputation (FHDI) method of Kim and Fuller (2004) is described. In Section 2.5, multiple imputation (MI) is described. A comparison of FHDI and MI is given in Section 2.6. Further fractional imputation methods are then discussed. Fractional regression hot deck imputation (FRHDI) is described in Section 2.7. Fractional nearest neighbor imputation (FNNI) is discussed in Section 2.8.

2.4.1. Kim and Fuller (2004): FHDI

Kim and Fuller (2004) studied fractional hot deck imputation (FHDI) under a cell mean model. In the cell mean model, one assumes that the finite population $U$ is made up of $G$ groups, which are used as imputation cells. Within cell $g$, the elements in the finite population are a realization of independent and identically distributed random variables with mean $\mu_g$ and variance $\sigma_g^2$. Then the distribution of elements in the imputation cell $g$ can be written for $i \in U_g$ as

$$y_i \sim ^{i.i.} (\mu_g, \sigma_g^2),$$

where $U_g$ is the set of indices for the $g^{th}$ imputation cell in the population $U$ and $\sim ^{i.i.}$ is an abbreviation for independent and identically distributed.

Under the assumption of an ignorable sampling mechanism and an ignorable response mechanism conditional on the cell, the above distribution of $y_i$ holds for both respondents and non-respondents alike. It is the case that

$$y_i \mid (A, A_r) \sim ^{i.i.} (\mu_g, \sigma_g^2), \quad \text{for every } i \in A_g,$$

where $A_g$ is the set of indices of sample elements in cell $g$ and $A_r$ is the set of indices of the sample respondents. Through the $y_i$ may be related to the design in the population, the division into imputation cells removes that dependence within the cell. Kim and Fuller (2004) label model (2.4.7) the cell mean model.
Let $A_{rg}$ be the indices for the respondents in cell $g$. Let $A_{mg}$ be the indices for the nonrespondents in cell $g$. Let $A_{Dj}$ be the indices for the donors for case $j$ from cell $g$, where $j$ is from $A_{mg}$ and $A_{Dj}$ is a subset of $A_{rg}$. Under the cell mean model, the linear hot deck imputed estimator of $\theta_{1N}$ can be constructed as

$$\hat{\theta}_{1n} = \sum_{g=1}^{G} \left( \sum_{i \in A_{rg}} w_i y_i + \sum_{j \in A_{mg}} w_j \sum_{i \in A_{Dj}} w_{ij}^* y_i \right).$$

(2.4.8)

This estimator is called a fractionally imputed estimator. For missing case $j$, $A_{Dj}$ is used as a donor set.

The fully efficient fractionally imputed (FEFI) estimator can be explained by the following description. Given nonresponse, one estimator of $\theta_{1N}$ is the ratio estimator

$$\hat{\theta}_{FE,1n} = \sum_{g=1}^{G} \left( \sum_{i \in A_{g}} w_i \right) \frac{\sum_{i \in A_{rg}} w_i y_i}{\sum_{i \in A_{rg}} w_i}. \tag{2.4.9}$$

It is called fully efficient as an imputation estimator because it contains no variability due to random selection of donors. The fractional hot deck imputed estimator can be constructed to be algebraically equal to the estimator (2.4.9) by choosing proper fractional weights $w_{ij}^*$. The fully efficient estimator (2.4.9) can be rearranged as

$$\hat{\theta}_{FE,1n} = \sum_{g=1}^{G} \left( \sum_{i \in A_{rg}} w_i \right) \frac{\sum_{i \in A_{rg}} w_i y_i}{\sum_{i \in A_{rg}} w_i}$$

$$= \sum_{g=1}^{G} \sum_{i \in A_{rg}} w_i \left( \frac{\sum_{i \in A_{rg}} w_i}{\sum_{i \in A_{rg}} w_i} \right) y_i$$

$$= \sum_{g=1}^{G} \sum_{i \in A_{rg}} w_i \left( \frac{\sum_{i \in A_{rg}} w_i + \sum_{i \in A_{mg}} w_i}{\sum_{i \in A_{rg}} w_i} \right) y_i$$

$$= \sum_{g=1}^{G} \left[ \sum_{i \in A_{rg}} w_i y_i + \sum_{j \in A_{mg}} w_j \sum_{i \in A_{rg}} \left( \frac{w_i}{\sum_{i \in A_{rg}} w_i} \right) y_i \right] \tag{2.4.10}$$

Therefore (2.4.10) gives a way to construct fractional weights when $A_{rg}$ is used as a donor set for each missing unit in the cell $g$. In this case, $A_{Dj} = A_{rg}$ for $j \in A_g$. 

implying that every observed unit in the imputation cell \( g \) is used as a donor. The resulting fractionally imputed estimator is

\[
\hat{\theta}_{1n}^{I} = \sum_{g=1}^{G} \left( \sum_{i \in A_{rg}} w_{iy_{i}} + \sum_{j \in A_{mg}} w_{j} \sum_{i \in A_{rg}} w_{ij}^{*} y_{i} \right),
\]  

(2.4.11)

where for \( j \in A_{mg} \) and \( i \in A_{rg} \)

\[
w_{ij}^{*} = \left( \sum_{i \in A_{rg}} w_{i} \right)^{-1} w_{i}.
\]  

(2.4.12)

Formula (2.4.11) is the generalization of (2.4.4) for the cell mean model. Then the estimator (2.4.11) is algebraically equivalent to (2.4.9) and is called the fully efficient fractionally imputed estimator. When the cell mean model is correct, the response mechanism is ignorable within cells, and imputation is done within cells, the linear hot deck imputed estimators (2.4.9) and (2.4.10) are unbiased. By (2.4.1) and (2.4.6),

\[
E(\hat{\theta}_{1n}^{I}|A, A_{r}, d) = E \left[ \sum_{g=1}^{G} \left( \sum_{i \in A_{rg}} w_{i} \mu_{g} + \sum_{j \in A_{mg}} w_{j} \sum_{i \in A_{Dj}} w_{ij}^{*} \mu_{g} \right) \right] |A, A_{r}, d,
\]  

(2.4.13)

When the complete sample estimators are unbiased under the sampling mechanism, the FHDI estimators are unbiased for \( \theta_{1N} \).

Now, we will focus on a more practical issue. Under a cell mean model, one does not have to assume a globally constant probability of nonresponse or a constant distribution across the whole sample. In practice, the imputation cells may be constructed by a function of observed auxiliary variables \( x \). Their basic assumption is that all elements in an imputation cell have the same probability of responding and the same probability distribution. It is reasonable idea if we can construct the imputation cells.

The FHDI estimator is unbiased if the assumed cell mean model is a true representation of the response mechanism. However, no matter what grouping we can construct
using the observed information, they are not likely to perfectly meet the requirement of equal response probability of all elements within an imputation cell. In addition, there is a chance to improve on the FHDI estimator since the information in auxiliary variables are only used to construct imputation cells in FHDI. To alleviate this problem, one possible approach is to construct a more fine set using auxiliary information so that the assumption of equal response probability is more likely to be true. One of motivations for the FNNI method of section 2.8 can be explained by this approach in that a nearest neighborhood constructed by a nearest neighbor technique is considered as a fine imputation cell.

2.4.2. Variance Estimation for FHDI

Under the cell mean assumptions discussed in Section 2.4.1, Kim and Fuller (2004) show that the variance of the hot deck imputation estimator (2.4.8) under the cell mean model is

\[
V(\hat{\theta}_{1n}) = V \left( \sum_{g=1}^{G} \sum_{i \in A_g} w_i \mu_g \right) + E \left( \sum_{g=1}^{G} \sum_{i \in A_g} w_i^2 \sigma_g^2 \right). \tag{2.4.14}
\]

The distribution used to define above variance is the joint distribution of the cell mean model, the sampling mechanism, the response mechanism, and the imputation mechanism or procedure.

To consider a variance estimator, they considered a replication variance estimator. For the complete sample, a replication variance estimator is

\[
\hat{V}(\hat{\theta}) = \sum_{k=1}^{K} c_k \left( \hat{\theta}^{(k)} - \hat{\theta} \right)^2, \tag{2.4.15}
\]

where \( \hat{\theta}^{(k)} \) is the \( k \)th estimate of \( \theta_N \), \( K \) is the number of replicates, and \( c_k \) is a factor associated with replicate \( k \) determined by the replication method. When a replication variance estimator (2.4.15) is used to estimate the variance of \( \hat{\theta}_{1n} \), the \( \hat{\theta}^{(k)}_{1n} \) can be written
\[ \hat{\theta}^{(k)} = \sum_{i \in A} w_i^{(k)} y_i, \]  

(2.4.16)

where \( w_i^{(k)} \) denotes the replicate weight for the \( i \)th unit of the \( k \)th replicate.

A naive way to apply the replication variance estimator to the completed data set (the data set with missing values filled in through imputation) is to treat the imputed values as real. When that is done, the naive replication variance estimator can be expressed as

\[ \hat{V}_{na} \left( \hat{\theta}^{I}_{1n} \right) = \sum_{k=1}^{K} c_k \left( \hat{\theta}^{I}_{na,1n} - \hat{\theta}^{I}_{1n} \right)^2, \]  

(2.4.17)

where

\[ \hat{\theta}^{I}_{na,1n} = \sum_{g=1}^{G} \left( \sum_{i \in A_{rg}} w_i^{(k)} y_i + \sum_{j \in A_{mg}} \sum_{i \in A_{Dj}} w_i^{(*)} y_i \right) =: \sum_{g=1}^{G} \sum_{i \in A_{rg}} w_i^{(*)} y_i. \]

When the complete sample variance estimator in (2.4.15) is design unbiased for the design variance of \( \theta_n \), the expectation of the naive variance estimator shown by Kim and Fuller (2004) is

\[ E \{ \hat{V}_{na} \left( \hat{\theta}^{I}_{1n} \right) \} = V \left( \sum_{g=1}^{G} \sum_{i \in A} w_i \mu_g \right) + E \left( \sum_{g=1}^{G} \sum_{i \in A_{rg}} \left[ \sum_{k=1}^{K} c_k (w_{na,i}^{(*)} - w_i^{(*)})^2 \right] \sigma_g^2 \right). \]  

(2.4.18)

Comparing (2.4.18) with (2.4.14), if \( w_{na,i}^{(*)} \) satisfies

\[ \sum_{i \in A_{rg}} \sum_{k=1}^{K} c_k (w_{na,i}^{(*)} - w_i^{(*)})^2 = \sum_{i \in A_{rg}} w_i^{(*)^2}, \quad g = 1, 2, \ldots, G \]  

(2.4.19)

then the naive variance estimator unbiasedly estimates the variance of the imputed estimator. Unfortunately, \( w_{na,i}^{(*)} \) constructed by the naive replication method do not
necessarily satisfy (2.4.20). This fact can be explained as follows: when the \(i^{th}\) unit is deleted on the \(k^{th}\) replicate, then \(w_i^{(k)}\) is equal to 0. The fractional weight \(w_i^*\), however, is not changed under the naive replication method. Consequently, the contribution of donor \(i\) to the total variance and its squared weight \(w_i^*\) is different from the contribution of donor \(i\) to the expectation of the naive variance estimator \((\sum_{k=1}^{K} c_k (w_{na,i}^* - w_i^*)^2)\). This indicates that \(w_{na,i}^*\) is not the proper weight to use in estimating the imputation variance.

To correct this problem, Kim and Fuller suggested the method of adjusting the fractional weight on each replicate so that the replicate weights can be defined as

\[
\hat{\theta}_{ad,1n}^{I(k)} = \sum_{g=1}^{G} \left( \sum_{i \in A_{rg}} w_i^{(k)} y_i + \sum_{j \in A_{mg}} w_j^{(k)} \sum_{i \in A_{Dj}} w_{ij}^* y_i \right)
\]

A natural starting place for modifying the fractional weight for donors is to reduce properly the fractional replicate weight \(w_{ij}^*\) without setting it equal to zero and increase properly the other \(M-1\) fractional replicate weights, \(\{w_{ij}^*, (t \in A_{Dj} \text{ and } t \neq i)\}\), when the \(i^{th}\) donor unit is deleted on the \(k^{th}\) replicate. This can be done in a way to construct an unbiased estimator of the variance. Theorem 2 (Kim and Fuller 2004) gives two conditions necessary to construct an adjusted replicate \(w_{ij}^*\) weight for unbiased variance estimation when the complete sample replication variance estimator of (2.4.15) is design-unbiased for the design variance of the complete sample estimator. Two conditions are

\[
\sum_{i \in A_{rg}} \sum_{k=1}^{K} c_k (w_{ad,i}^* - w_i^*)^2 = \sum_{i \in A_{rg}} w_i^{*2}, \quad g = 1, 2, \ldots, G, \quad (2.4.20)
\]

and

\[
\sum_{i \in A_{r}} w_{ad,i}^* = 1 \quad \text{for all } j \in A. \quad (2.4.21)
\]

There are numerous ways of constructing replicates that satisfy (2.4.20) and (2.4.21). Following Kim and Fuller (2004), we present their proposed method of finding the proper
amount of adjustment to obtain the desired expectation. For simplicity of notation, assume there is only one imputation cell and $G = 1$.

First, compute replication weight in the ordinary way.

$$w_i^{(k)} = \begin{cases} w_i \frac{\sum_{A^{(k)}} w_i}{\sum_{A^{(k)}} w_i} & \text{if } i \neq k \\ 0 & \text{if } k = i, \end{cases}$$

(2.4.22)

where $A^{(k)}$ is the $k^{\text{th}}$ replicate sample.

Second, compute the naive replicate weights,

$$w_{na,i}^{(k)} = w_i^{(k)} + \sum_{i \in A_mg} w_j^{(k)} w_{ij}^*.$$

Third, for a fractional imputation procedure with $M$ distinct donors, define replicates for missing unit $j$ of replicate $k$ as

$$w_{ij}^{*(k)} = \begin{cases} w_{ij}^* - b_k & \text{if } k \in A_{Dj} \text{ and } k = i \\ w_{ij}^* + (M - 1)^{-1}b_k & \text{if } k \in A_{Dj} \text{ and } k \neq i \\ w_{ij}^* & \text{otherwise}. \end{cases}$$

(2.4.23)

The replicates defined in (2.4.23) satisfy (2.4.21) for any value of $b_k$. Fourth, form a quadratic function involving $b_k$ and solve for $b_k$ to satisfy condition (2.4.20). Kim and Fuller (2004) give the equation to be solved. Doing so gives a value of $b_k$, values of $w_{ij}^{*(k)}$, and values of

$$w_{ad,i}^{(k)} = w_i^{(k)} + \sum_{i \in A_mg} w_j^{(k)} w_{ij}^{*(k)},$$

which result in unbiased variance estimator.

2.4.3. Inference for FHDI

Statistical inference for parameters based on FHDI is based on central limit theorem arguments. Suppose $\hat{\theta}$ is the estimate of $\theta$ and $\hat{V}$ is the estimate of variance. A confidence
interval for $\theta$ is constructed based on the normal distribution: $(\hat{\theta} - z^*\sqrt{V}, \hat{\theta} + z^*\sqrt{V})$, where $z^*$ is a critical value from a standard normal distribution corresponding to the desired level of confidence. This method should apply to other fractional imputation methods as well. If $\theta_0$ is a null value for $\theta$, then hypotheses tests can be based on comparing $(\hat{\theta} - \theta_0)/\sqrt{V}$ to the standard normal distribution. For two-sided tests, values with absolute value above the $z^*$ critical value suggest that the null hypothesis should be rejected. Note that if $w_i$ is not equal within imputation cell, (2.4.9) is no longer fully efficient under the cell mean model. An approximation to the fully efficient procedure where $A_{Dj}$ of relatively small size is used as donor sets are proposed by Fuller and Kim (2005a). Their resulting estimator is a fully efficient estimator of the mean. Estimates of the cumulative distribution are nearly fully efficient.

2.5. Multiple Imputation: MI

Another popular imputation method in practice is multiple imputation (MI). Multiple imputation, proposed by Rubin (1978), is a procedure for handling missing data that allows the data analyst to use analyses designed for a complete data set while at the same time providing a method for estimating the uncertainty due to the missing data. The primary goal of multiple imputation is to simplify estimation and to provide a valid estimator of the variance.

In case of general multiple imputation, repeated imputations are drawn from the posterior predictive distribution of the missing values. The posterior predictive distribution is determined by the prior distribution on model parameters, the specified data model, and the observe data. Multiple imputation creates $M$ different imputed data sets, one for each complete set of imputations for all the missing values. For each completed data set, one computes the desired analysis and produces both an estimate of the quantity of interest and a variance estimate for the estimator. The several estimates and variance
estimates are then combined according to specified formulas.

Let $\hat{\theta}_{I(k)}$ be the standard estimate of parameter $\theta$. Let $\hat{V}_{I(k)}$ be the standard variance estimator that treats the imputed values as if they were observed based on the $k^{th}$ completed data set. Then the multiple imputation point estimator of $\theta$ is

$$\hat{\theta}_M = \frac{1}{M} \sum_{k=1}^{M} \hat{\theta}_{I(k)}. \quad (2.5.1)$$

The associated variance estimator is

$$\hat{V}_M = W_{M} + (1 + \frac{1}{M})B_{M}, \quad (2.5.2)$$

where

$$W_{M} = M^{-1} \sum_{k=1}^{M} \hat{V}_{I(k)},$$

and

$$B_{M} = (M - 1)^{-1} \sum_{k=1}^{M} (\hat{\theta}_{I(k)} - \hat{\theta}_M)^2.$$ 

The term $W_{M}$ is called the average within imputation data set variance. The term $B_{M}$ expresses the variability of estimates across sets of imputations. The term $1 + 1/M$ is greater than 1 to account for the fact that a finite number of imputations ($M$) are used for each missing value. An infinite number of imputations theoretically would be needed to express the posterior predictive distribution of a missing value.

Main advantages of multiple imputation are to allow the data analyst to use standard techniques of analysis designed for complete data and simultaneously provide a variance estimator easy to use. In the literature, however, several authors have pointed out some limitations to the multiple imputation approach. These authors include Fay (1992, 1993), Meng (1994), Kott (1995), Kim (2002), and Kim and Fuller (2004). The primary concern of Fay, Meng, and Kott is when the model used by the person doing the imputation is
not the same as the model used by the person analyzing the data. In particular, if the imputation model (and associated prior distribution) do not include model elements that are present in the analysis model (such as interactions between variables and some predictor variables), then the analyst will not necessarily be able to produce reasonable estimates of the associated parameters. Kim and Fuller concern themselves with whether the multiple imputation estimator of variance is simply too large. They point out that if one is interested in a subset of the population and the multiple imputation procedure uses cases from beyond this subset in modeling, then extra variation can be introduced into the imputations on the subset of interest. More detail will be presented on next section.

A particular implementation of multiple imputation, commonly referred to as the approximate Bayesian bootstrap (described in Rubin and Schenker 1986) can be viewed as a hot deck imputation method in the multiple imputation context. Under the cell mean model (2.4.6), the approximate Bayesian bootstrap imputation procedure can be described as follows. Repeat the following $M$ times, $m = 1, \ldots, M$.

Step 1. A donor set of size $r_g$ for cell $g$ is selected with replacement with equal probabilities of selection from the respondents in each cell. This is done for all cells, $g = 1, \ldots, G$.

Step 2. For each missing unit $j \in A_m$, imputed value $y^{*}_{j(m)}$ is drawn from the donor set produced in Step 1 with replacement and with equal probabilities of selection.

Now use the multiple imputation estimator and variance estimator formulas as described above.

For large sample sizes and small $B$, the reference distribution for interval estimates and significance tests is to replace the normal reference distribution by a $t$ distribution,

$$
\frac{\hat{\theta}_M - \theta}{\sqrt{V_M}} \sim t_v, \quad (2.5.3)
$$
with degrees of freedom given by

\[ v = (B - 1) \left( 1 + \frac{1}{B + 1} \frac{W_M}{B_M} \right)^2, \]  

(2.5.4)

based on a Satterthwaite approximation. An improved expression for the degree of freedom for small sample sizes and small \( B \) is

\[ v^* = (v^{-1} + \hat{v}_{obs}^{-1})^{-1}, \]  

(2.5.5)

where

\[ \hat{v}_{obs}^{-1} = (1 - \hat{\gamma}_B) \left( \frac{v_{com} + 1}{v_{com} + 3} \right) v_{com}, \]

\[ \hat{\gamma}_B = 1 - \frac{W_M}{V_m}, \]

estimates the fraction of missing information and \( v_{com} \) is the degree of freedom for the completed data set. The theoretical basis for (2.5.4) is given in Rubin and Schenker (1986), and for (2.5.5) is given in Barnard and Rubin (1999).

Multiple imputation (MI) produces overall estimates of parameters and corresponding sampling variances. Li, Raghunathan, and Rubin (1991) introduced a modified Wald test procedure for testing hypotheses. In order to use the Wald test procedure, it is necessary to create many more imputations than the dimension of the parameters (see, for example, Meng and Rubin (1992, page 105). Meng and Rubin (1992) introduced a likelihood ratio based procedure for testing with multiply imputed data. Li, Meng, Raghunathan, and Rubin (1991) introduced a procedure for combining results from either Wald test statistics or LRT statistics. Schafer (1997) also describes these procedures. Applications in this dissertation focus on point estimations, estimating variances, and confidence intervals, so hypothesis testing methods are not studied further.
2.6. Comparison of Fractional Hot Deck Imputation (FHDI) and MI

Kim and Fuller (2004) compared fractional hot deck imputation under the cell mean model to multiple imputation (MI; Rubin 1987) using a corresponding likelihood model and standard non-informative Bayesian prior distribution on unknown parameters. In their setting fractional imputation was more efficient for estimating the overall mean than multiple imputation with the same number of donors. In addition, their variance estimator was less variable than the variance estimator corresponding to multiple imputation. Their variance estimates with FEFI had smaller bias and much smaller variance than the multiple imputation variance estimator with the same number of imputations. The improvement can be explained by the fact that MI, by drawing parameters from a posterior distribution in addition to drawing imputed values from a data model, adds additional variability to the variance of the imputation-based estimator. The work of Kim and Fuller (2004) also demonstrates that if a domain of interest is not included in a multiple imputation model that multiple imputation can produce quite poor estimates. This occurs because the between-domain variability becomes included in the unit-level variability during generation of imputation values from the posterior predictive distribution. Of course, one would really want to use domains of interest in the multiple imputation model, but that can be challenging to do if there are potentially many variables defining domains. Fractional imputation only can use cases that are observed to be members of a domain in order to estimate the characteristics of the domain. The unequal changes in weights accompanying the FI procedure could lead to some increase in variance for estimation in domain, but in the situations compared these increases seem small in comparison to between-domain variation.

The fractional imputation procedure produces an easy-to-use data set that is a rectangular data array containing a set of characteristics for respondents, associated weights for point estimation, and replicate weights for variance estimation. The same replicate
weights can be used for variance estimation for imputed variables, variables observed on all respondents, and under some model assumptions, for functions of the two types of variables. In addition, variance estimation for other variables such domain characteristics is easy to implement using the replicates with much smaller biases than those of multiple imputation (Kim and Fuller 2004). The imputed data set generated by FI method may be appropriate for both planned and unplanned estimates.

2.7. Fractional Regression Hot Deck Imputation: FRHDI

One way to preserve the correlation structure under hot deck imputation using fractional imputation is fractional regression hot deck imputation (FRHDI). FRHDI, suggested by Kim (2007), imputes multiple values for each instance of a missing dependent variable. The imputed values are equal to the predicted value based on the fully observed cases plus multiple random residuals chosen from the set of empirical residuals. Fractional weights are chosen to enable variance estimation and to preserve the correlation among independent and dependent variables. To apply regression imputation to fractional imputation, the weighted mean of the imputed values using stochastic regression imputation is used to impute the missing data. By modifying fractional weights, the fractional regression imputation can take the form of hot deck fractional imputation. As a result, FRHDI preserves the correlation structure and uses observed values for imputation like hot deck imputation.

In this section, we briefly summarize the procedure of FRHDI and explain the limitations of FRHDI. First, the imputed value by stochastic regression imputation under fractional imputation is

$$y_{ij}^* = \sum_{i \in A_{Dj}} w_{ij}^*(\hat{y}_i + \hat{e}_i),$$

(2.7.1)

where $A_{Dj}$ is the set of donors for $j \in A_m$, $A_{Dj} = \{i; d_{ij} = 1\}$. The imputed estimator
of $\theta_1N$ and $\theta_2N$ under fractional imputation can be constructed as follows:

$$\hat{\theta}_{1n}^I = \sum_{i \in A_r} w_i y_i + \sum_{j \in A_m} w_j y_{ij}^*$$

$$= \sum_{i \in A_r} w_i y_i + \sum_{j \in A_m} \sum_{i \in A_D} w_j w_{ij}^* (\hat{y}_j + \hat{e}_i)$$

and

$$\hat{\theta}_{2n}^I = \sum_{i \in A_r} w_i x_i y_i + \sum_{j \in A_m} \sum_{i \in A_D} w_j w_{ij}^* x_j (\hat{y}_j + \hat{e}_i).$$

When $\hat{y}_j$ is a linear function of $x_j$, then the weighted mean of the imputed values can be written as

$$\sum_{i \in A_D} w_{ij}^* (\hat{y}_j + \hat{e}_i) = \sum_{i \in A_D} w_{ij}^* (y_i + (x_j - x_i) \hat{\beta}_1)$$

$$= \sum_{i \in A_D} w_{ij}^* y_i + \hat{\beta}_1 \sum_{i \in A_D} w_{ij}^* (x_j - x_i). \quad (2.7.2)$$

The second term on the right side of (2.7.2) can be zero by using adjusted fractional weights. The sufficient and necessary condition for taking the form of the fractional hot deck imputation (FHDI), i.e. of having the second term be zero, is

$$\sum_{i \in A_D} w_{ij}^* (1, x_i) = (1, x_j) \quad \text{for } j \in A_m. \quad (2.7.3)$$

The FRHDI estimator can be expressed in the form of the fractional hot deck imputed estimator as follows:

$$\hat{\theta}_{1n}^I = \sum_{i \in A_r} w_i y_i + \sum_{j \in A_m} \sum_{i \in A_D} w_{ij}^* y_i \quad \text{(2.7.4)}$$

and

$$\hat{\theta}_{2n}^I = \sum_{i \in A_r} w_i x_i y_i + \sum_{j \in A_m} \sum_{i \in A_D} w_{ij}^* x_i y_i, \quad \text{(2.7.5)}$$

where the $w_{ij}^*$ satisfy condition (2.7.3).
Fractional weights satisfying condition (2.7.3) can be constructed using the regression technique. The modified fractional weight is

\[ w_{ij}^* = \alpha_{ij} + (x_j - \bar{x}_I)S_{xx,j}^{-1}\alpha_{ij}(x_i - \bar{x}_I) \]  

(2.7.6)

where

\[ S_{xx,j} = \sum_{i \in A_{Dj}} \alpha_{ij}(x_i - \bar{x}_I)^2, \]

\[ \bar{x}_I = \sum_{i \in A_{Dj}} \alpha_{ij}x_i, \]

\[ \alpha_{ij} = 1/M, \]

and \( M \) is the number of donors used for fractional imputation.

To show the unbiasness of the FRHDI estimator, Kim (2007) used the cell regression model. Assume that the sample \( A \) is made up of \( G \) imputation cells and the imputed residuals are selected in the same cell. The cell regression model is

\[ E(y_i | x_i, i \in A_g) = \beta_0 + \beta_1 x_i \]

\[ V(y_i | x_i, i \in A_g) = \sigma_g^2. \]  

(2.7.7)

Under the cell regression model and ignorable response sampling regression mechanisms, the FRHDI estimators of \( \theta_{1N} \) and \( \theta_{2N} \) are conditionally unbiased. Since the complete sample estimators are unbiased under the sampling mechanism, FRHDI estimators are unbiased for both parameters.

For variance estimation of the FRHDI estimator, Kim (2007) modified a replication variance estimator by Kim and Fuller (2004). They change the fractional replicate weights so that the expected value of the sum of squares is changed by the proper amount given that they have replaced the cell mean model by the cell regression model. Under FRHDI, two conditions are needed for the replicated fractional weights to achieve unbiasedness. The two conditions are

\[ \sum_{i \in A_{Dj}} \sum_{k=1}^{K} c_k (w_{adj,i}^{(k)} - w_i^*)^2 = \sum_{i \in A_{Dj}} w_i^{*2}, \quad g = 1, 2, \ldots, G, \]  

(2.7.8)
and

\[ \sum_{i \in A_{Dj}} w_{adj,i}^{*(k)}(1, x_i) = (1, x_j) \quad \text{for all } j \in A, \tag{2.7.9} \]

where

\[ w_{adj,i}^{*(k)} = w_i^{(k)} + \sum_{j \in A_m} w_j^{(k)} w_{ij}^{*(k)}. \]

Note that under the cell mean model with \( x_i \equiv 0 \), these two conditions are the same conditions discussed in Section 2.4.2. The replicated fractional weight satisfying (2.7.9) in replicate \( k \) for the value donated by \( i \) to \( j \) can be constructed by

\[
 w_{ij}^{*(k)} = \begin{cases} 
 w_{ij}^{*} - (1 - \alpha_{ij})b_k + h_{ik,j}b_k & \text{if } k \in A_{Dj} \text{ and } k = i \\
 w_{ij}^{*} + \alpha_{ij}b_k + h_{ik,j}b_k & \text{if } k \in A_{Dj} \text{ and } k \neq i \\
 w_{ij}^{*} & \text{otherwise,} \end{cases} \tag{2.7.10}
\]

where

\[ h_{ik,j} = \alpha_{ij}(x_i - \bar{x}_{Ij})S_{xx,j}^{-1}(x_k - \bar{x}_{Ij}) \]

and \( b_k \) is to be determined. Under this set up, the replicated fractional weights satisfy (2.7.9) for any value of \( b_k \). Kim (2007) forms a quadratic function from which they can solve for \( b_k \) in the same way of Kim and Fuller (2004). Then \( b_k \) can be easily determined by the quadratic formula and the resulting variance estimator will be unbiased.

The proposed procedure by Kim (2007) can be motivated as follows. First, although the procedure creates imputations (a prediction plus a random residual) that might not equal actually observed values, the resulting estimator is equivalent to one using only observed values with adjusted weights as is the case for hot deck imputation. Second, it is easy to estimate the variance of the imputed estimator by applying a consistent replication variance estimation procedure as with the fractional imputation procedure suggested by Kim and Fuller (2004). Third, FRHDI is expected to be more robust against model violations in that FRHDI uses fractional weights constructed using the
relationship between the auxiliary variable \( x_i \) in \( A_r \) and \( x_j \) in \( A_m \), which does not depend on the estimate of \( \beta \).

There is some limitation to FRHDI. When the FRHDI method is applied to multivariate variables, it may not possible to find stable fractional regression weights satisfying condition (2.7.3). That is, the fractional weights constructed by the regression weighting technique in FRHDI can be variable: the procedure can produce some large weights, or some that are negative. A large weight on donors can result in large imputation variance for some estimators especially when a small number of donors is chosen. In addition, a negative fractional weight can be seriously problematic for estimating the variance of imputed estimators since \( b_k \) may not be determined by solving a quadratic function of \( b_k \).

2.8. Fractional Nearest Neighbor Imputation: FNNI

Nearest neighbor imputation (NNI) selects the respondent closest to the nonrespondent and inserts the respondent value for the missing item. If the relationship between \( y \) and \( x \) is linear and distributions of \( x \) are symmetric for respondents and nonrespondents then the NNI estimator is unbiased. When the distribution of \( x \) is not symmetric, the NNI estimator using single imputation has bias of order \( r^{-1} \) where \( r \) is the size of \( A_r \).

Instead of using just one nearest neighbor as a donor, one can consider using multiple nearest neighbor donors for each case with a missing measurement. Using multiple donors could improve the efficiency of the point estimator by averaging over multiple closer neighbors, thereby reducing imputation variance, and be more robust than single nearest neighbor imputation in terms of the assumptions on the distribution of \( x \). Fuller and Kim (2005b) studied the implementation of replicated nearest neighbor imputation under fractional imputation. In their study, variance estimation methodology suggested by Kim and Fuller (2004) is modified for a variance estimator under the NNI method without the imputation classes constructed using pres existing strata. They assumed
that an adequate approximation for the distribution of elements in the neighborhood is

\[ y_i \sim i.i.d. \left( \mu_j, \sigma_j^2 \right) \quad i \in B_j \quad (2.8.1) \]

where

\[ \mu_j = E\{y_j | j \in B_j\} \]
\[ \sigma_j^2 = E\{(y_j - \mu_j)^2 | j \in B_j\}, \]

and \( B_j \) is the set of indices for the elements in the neighborhood of element \( j \). Under the model (2.8.1) and an ignorable response mechanism, the FNNI estimators of \( \theta_{1N} \) and \( \theta_{2N} \) are unbiased. Since the validity of the Kim and Fuller method described in Section 2.4.2 does not require the cells to be disjoint, the variance estimation procedure in Section 2.4.2 can be applied to estimate the variance of the FNNI estimator replacing the cell mean model by the nearest neighbor cell model (2.8.1).

3. Some theoretical results for imputation methods

In the previous section, we presented various imputation procedures. In this section, we study some theoretical results for Regression Imputation, Fractional Hot Deck Imputation and Fractional Regression Hot Deck imputation.

3.1. Basic Setup

Let \( U = \{1, 2, \ldots, N\} \) be a set of indices of a finite population \( F_N \) of \( N \) elements with \( N \) known. Associated with the \( i^{\text{th}} \) unit of \( U \) is \( y_i \), a variable of interest, and \( x_i = \{1, x_{i1}, \ldots, x_{ip}\}' \), an auxiliary variable (or vector of variables), where some \( y_i \) are missing but every \( x_i \) is observed. Let \( A \) denote the set of indices of the elements in a sample of size \( n \) selected by the chosen sampling mechanism.

In limit theorems, we imagine a sequence of populations and samples increasing in size. Let the populations be indexed by \( t = 1, 2, \ldots \) such that the populations are nested:
Then the populations sizes are increasing: \( N_1 < N_2 < \ldots \). It is assumed that the sample sizes are increasing \( (n_1 < n_2 < \ldots) \), but the samples are not necessarily nested.

Let the population quantity of interest be \( Y_N = \sum_{i=1}^{N} y_i \). Let \( \hat{Y}_n \) be an estimator of \( Y_N \) based on complete response from sample members,

\[
\hat{Y}_n = \sum_{i \in A} w_i y_i \tag{3.1}
\]

where \( w_i = 1/\pi_i \) is a sampling weight for unit \( i \) that depends on the sampling mechanism and \( \pi_i = Pr(i \in A) \) is the inclusion probability. A \textit{design unbiased} estimator is defined in the following definition.

**Definition 1.** A statistic \( \hat{\theta}_n \) is said to be design unbiased for the finite population parameter \( \theta_N \) if

\[
E\{\hat{\theta}_n|F_N\} = \theta_N,
\]

where conditioning on \( F_N \) denotes expectation with respect to the sampling mechanism in the realized finite population. The expectation above is taken with respect to the sampling mechanism, which is generated by repeated application of the sample selection method. The estimator \( \hat{Y}_n \) is design unbiased for the population total \( Y_N \).

A design unbiased variance estimator is given by the following formula (Särndal, Swensson and Wretman, 1992):

\[
\hat{V}_n = \sum_{i \in A} \sum_{j \in A} \Delta_{ij} w_i y_i w_j y_j, \tag{3.2}
\]

where \( \Delta_{ij} = (\pi_{ij} - \pi_i \pi_j)/\pi_{ij} \) and \( \pi_{ij} = Pr(i, j \in A) \) is the joint inclusion probability.

Under nonresponse, we define the response indicator variable for \( y_i \) by

\[
R_i = \begin{cases} 
1 & \text{if } y_i \text{ responds}, \\
0 & \text{if } y_i \text{ does not respond}.
\end{cases}
\]
Imputation methods generally require assumptions of two types in order to be appropriate for use with a particular study variable. One assumption is expressed by the imputation model and the other is expressed by the response model. The imputation model makes an assumption about the distribution of the study variable and its relationship with other variables collected in the survey. The other model is called the response model and concerns assumptions about the probability of obtaining respondents from the sample.

The outline of the work in this section is as follows. Some order concepts useful in investigating the asymptotic properties of the imputed estimator are described. Some summaries of properties of imputed estimators follow.

3.2. Order in probability

Order of magnitude is useful for describing asymptotic properties of estimators. The concept of order of convergence is used in proofs. In this section, some concepts of order in probability are summarized from Fuller (1996, section 5.1). Let \( a_n \) be a sequence of real numbers. Let \( g_n \) be a sequence of positive real numbers. Let \( Y_n \) be a sequence of random variables.

**Definition 2.** We say \( a_n \) is of smaller order than \( g_n \) and write

\[
a_n = o(g_n)
\]

if

\[
\lim_{n \to \infty} g_n^{-1} a_n = 0.
\]

**Definition 3.** We say \( a_n \) is at most of order \( g_n \) and write

\[
a_n = O(g_n)
\]

if there exists a real number \( M \) such that \( g_n^{-1}|a_n| \leq M \) for all \( n \).
For example, if \( a_n = \sqrt{n} \) and \( g_n = n \), then \( a_n/g_n = 1/\sqrt{n} \) and \( a_n = o(g_n) \) because \( \lim 1/\sqrt{n} = 0 \). It also is true that the maximum value of \( a_n/g_n \) occurs for \( n = 1 \) when the value of the ratio is 1. Therefore, \( a_n = O(g_n) \) as well, because \( a_n/g_n \leq 1 \) for all \( n \). One concept of order does not imply the other. For example, if \( a_n = g_n \), then \( a_n = O(g_n) \) but \( a_n \neq o(g_n) \). Conversely, if \( a_n = \sqrt{n} \) and \( g_n = n - 1 \), then \( a_n = o(g_n) \) but \( a_n/g_n \) is infinite at \( n = 1 \) so that \( a_n \neq O(g_n) \).

The concepts of order for random variables, introduced by Mann and Wald (1943), are closely related to convergence in probability. Below is a definition of convergence in probability of random variables.

**Definition 4.** The sequence of random variables \( \{Y_n\} \) converges in probability to the random variable \( Y \), and we write

\[
Y_n \overset{P}{\rightarrow} Y
\]

if for every \( \epsilon > 0 \)

\[
\lim_{n \to \infty} P\{|Y_n - Y| > \epsilon\} = 0.
\]

An equivalent definition is that for every \( \epsilon > 0 \) and \( \delta < 0 \) there exists an \( n_\epsilon \) such that for all \( n > n_\epsilon \),

\[
P\{|Y_n - Y| > \epsilon\} < \delta.
\]

In the definition above, if random variable \( Y \) is a constant random variable, i.e., \( Y \) has one value with probability one, then the sequence of random variables \( Y_n \) converges in probability to the constant. For example, if \( Y_n \) is the mean of \( n \) iid random variables with mean \( \mu \) and variance \( \sigma^2 \), then \( Y_n \) converges in probability to \( \mu \). In that case, one could write \( \lim_{n \to \infty} P(|Y_n - \mu| > \epsilon) = 0 \) and be clear.

Convergence in probability can have associated orders.

**Definition 5.** We say \( Y_n \) is of smaller order in probability than \( g_n \) and write

\[
Y_n = o_P(g_n)
\]
if

\[ g_n^{-1} Y_n \overset{p}{\to} 0 \]

That is, \( \lim_{n \to \infty} P(|Y_n/g_n - 0| > \epsilon) = 0 \).

**Definition 6.** We say \( Y_n \) is at most of order in probability \( g_n \) (or bounded in probability \( g_n \)) and write

\[ Y_n = O_p(g_n) \]

if, for every \( \epsilon > 0 \), there exists a positive number \( M_\epsilon \) such that

\[ P\{|Y_n| > M_\epsilon g_n\} \leq \epsilon \]

for all \( n \).

Analogous definitions hold for vectors as described in Fuller (1996, section 5.1).

Any random variable with finite variance is bounded in probability by the square root of its second moment about the origin by Chebyshev’s inequality. Chebyshev’s inequality is stated and proved as Theorem 5.1.1 of Fuller (1996). The statement two sentences before follows from this result with \( r = 2 \).

**Theorem 1.** Let \( r > 0 \). Let \( Y \) be a random variable with \( E\{|Y|^r\} < \infty \). Let \( F(x) \) be the distribution function of \( X \). For every \( \epsilon > 0 \) and finite value \( a \),

\[ P\{|Y - a| \geq \epsilon\} \leq E\{|Y - a|^r\}/\epsilon^r \]

Corollary 5.1.1.1 of Fuller (1996) uses Chebychev’s result and the ideas of orders of convergence to relate convergence in expectation of a squared random variable to convergence in probability. The result is stated below.

**Corollary 1.** Let \( \{Y_n\} \) be a sequence of random variables and \( g_n \) a sequence of positive real numbers such that

\[ E\{Y_n^2\} = O(g_n^2) \]

Then
\[ Y_n = O_p(g_n). \]

**Proof.** By assumption, there exists an \( M \) such that \( E\{Y_n^2\} < M^2 g_n^2 \) for all \( n \). By Chebyshev’s inequality, letting \( a = 0 \) and \( r = 1 \), for every \( \epsilon > 0 \), there exists a positive \( M_\epsilon \geq M \epsilon^{-0.5} \) such that

\[
P(|Y_n| \geq g_n M_\epsilon) \leq \frac{E\{Y_n^2\}}{g_n^2 M_\epsilon^2} < \frac{M^2 g_n^2}{g_n^2 M_\epsilon^2} < \epsilon.
\]

\(\Box\)

### 3.3. Results for the Linear Regression Imputation Estimator

In the literature, imputation methods using an assumed imputation model are sometimes called population model approaches. Under the population model approach, the study variable \( y \) is assumed to be generated by some stochastic mechanism. This stochastic mechanism is referred to as the superpopulation model. A linear regression model for the superpopulation model is as follows:

\[
E(y_i|x_i) = x_i^\prime \beta_0, \tag{3.3.1}
\]

\[
V(y_i|x_i) = \sigma^2,
\]

and

\[
cov(y_i, y_j|x_i) = 0 \text{ if } i \neq j
\]

for \((p+1)\)-dimensional vector \( \beta_0 \). In this situation, we may say that the finite population is a sample of size \( N \) from the superpopulation model. Under the population model approach, the distribution of the \( R_i \) does not need to be specified. Deville and Särndal (1994) and Rao (1996) discussed the properties of the regression imputation estimator based on the population model approach. We assume that the response mechanism is
missing at random (MAR). That is, the distribution of $R_i$ depends only on $x_i$. Under linear regression imputation, $y_i^* = x_i'\hat{\beta}$ is used as an imputed value for missing $y_i$, where $\hat{\beta}$ is the solution of

$$U(\beta) \equiv \sum_{i \in A} w_i R_i x_i (y_i - x_i' \beta) = 0.$$  \hfill (3.3.2)

Then, the estimator of $Y$ based on the imputed values can be written as

$$\hat{Y}_I = \sum_{i \in A} w_i \{ R_i y_i + (1 - R_i) x_i' \hat{\beta} \}.$$  \hfill (3.3.3)

Under this set up, the conditional expectation of the regression imputation estimator $\hat{Y}_I$ of $Y_N$ is

$$E(\hat{Y}_I - \hat{Y}_n | X, R, A) = E \left( \sum_{i \in A} w_i (1 - R_i)(x_i\hat{\beta} - y_i) | X, R, A \right)$$  \hfill (3.3.4)

which is equal to zero since $E \left( (x_i\hat{\beta} - y_i) | X, R, A \right) = 0$, where $X = \{(i, x_i) : i \in A\}$, $R = \{(i, R_i) : i \in A\}$. In words, given any $A$, $X$, and $R$, the difference between the imputed estimator $\hat{Y}_I$ and the complete sample estimator $\hat{Y}_n$ differs from zero only by random error having zero expectation under the model. Therefore, the regression imputation estimator is unbiased since the complete sample estimator is unbiased under the sampling mechanism.

Deville and Särndal’s (1994) variance estimation method is developed by decomposing the variance of the imputed estimator. The variance of the imputed estimator is given by

$$V(\hat{Y}_I) = E \left( \hat{Y}_I - Y_N \right)^2 = E \left( \hat{Y}_I - \hat{Y}_n + \hat{Y}_n - Y_N \right)^2 = EV(\hat{Y}_n | F_N) + E \left( E \{(\hat{Y}_I - Y_N)^2 + 2(\hat{Y}_I - \hat{Y}_n)(\hat{Y}_n - Y_N) | X, R, A \} \right).$$

The reference distribution is the joint distribution of the superpopulation model, the sampling mechanism and the response mechanism. The reference distribution in the
conditional variance given $F_N$ is the sampling mechanism treating $F_N$ as fixed. The conditional distribution in the second half of the third line is given the realized sample, the realized $x$-values, and the realized respondents. The first term in the last equation is called the sampling variance in terms of variance for complete cases. The second term, called imputation variance, is the variance added through imputation. To estimate $V(\hat{Y}_I)$, Deville and Särndal consider estimating the two terms separately. If we know $y_i$ throughout the sample, the sampling variance could be estimated by using the design unbiased estimator $\hat{V}_n$ in (3.2). Since some $y_i$ values are missing in the sample, define the difference $C = E\{\hat{V}_n(\hat{Y}_n) - \hat{V}_n(\hat{Y}_I)|X, R, A\}$ and estimate it by $\hat{C}$ satisfying $E(\hat{C}|X, R, A) = C$. Then, the sampling variance can be unbiasedly estimated by $\hat{V}_n(\hat{Y}_I|F_N) + \hat{C}$. The imputation variance also can be unbiasedly estimated by $\hat{V}_{imp}$ satisfying $E(\hat{V}_{imp}|X, R, A)$ equal to the conditional expectation of the imputation variance.

Rao (1996) discussed the jackknife variance estimator based on the adjusted imputed values in detail for various imputation methods including regression imputation methods. The adjusted jackknife is approximately unbiased under the joint distribution of the superpopulation model, the sampling mechanism, and the response mechanism.

3.4. Results Concerning Fractional Imputation Estimators

3.4.1. Fractional Hot Deck Imputation

Kim and Fuller (2004) studied the use of fractional imputation under the population model approach. One assumes that the finite population $U$ is made up of $G$ groups, which are used as imputation cells. Imputation cells are subsets of the sample to which respondents are assigned based on characteristics. Within cell $g$, the elements in the finite population are a realization of independent and identically distributed random variables with mean $\mu_g$ and variance $\sigma^2_g$. Then the distribution of elements in the imputation cell
$g$ can be written for $i \in U_g$ as

$$y_i \sim^{i.i.} (\mu_g, \sigma^2_g),$$

where $U_g$ is the set of indices for the $g^{th}$ imputation cell in the population $U$ and $\sim^{i.i.}$ is an abbreviation for independent and identically distributed.

Under the assumption of an ignorable sampling mechanism and an ignorable response mechanism conditional on the cell, the above distribution of $y_i$ holds for both respondents and non-respondents alike. It is the case that

$$y_i| (A, A_r) \sim^{i.i.} (\mu_g, \sigma^2_g), \quad \text{for every } i \in A_g,$$

where $A_g$ is the set of indices of sample elements in cell $g$ and $A_r$ is the set of indices of the sample respondents. Though the $y_i$ may be related to the design in the population, the division into imputation cells removes that dependence within the cell. Kim and Fuller (2004) label model (3.4.2) the cell mean model.

Kim and Fuller (2004) describe fractional hot deck imputation and introduce a weighting system for this method. Define $d = \{d_{i,j}; i \in A_r, j \in A_m\}$ where $d_{i,j}$ is the number of times that $y_i$ is used as a donor for missing $y_j$. The distribution of $d$ is called the imputation mechanism. Define $w^*_{ij}$ to be the fraction of the original weight for element $j$ which donor $i$ donates for the missing $y_j$. Note that $w^*_{ii} = 1$ for $i \in A_r$ and $w^*_{ii} = 0$ for $i \in A_m$. The $w^*_{ij}$ is called the imputation fraction. The terminology is appropriate because it is the weight fraction that observed element $i$ donates to missing unit $j$. It is worth noting that imputation fractions necessarily sum to one. A simple linear estimator using hot deck imputation can now be written

$$\hat{Y}_{I,H} = \sum_{i \in A_r} \left( \sum_{j \in A_m} w_j w^*_{ij} \right) y_i = \sum_{i \in A_r} \alpha_i y_i,$$

where $\alpha_i$ is the total weight of donor $i$. If observed unit $i$ is never used as a donor, clearly $\alpha_i = w_i$. 
In their Theorem 1, the properties of the estimator (3.4.3) are presented. Under some conditions including an ignorable imputation mechanism, the estimator satisfies

\[ E(\hat{Y}_{I,H} - Y_N) = 0 \]  

and

\[ V(\hat{Y}_{I,H} - Y_N) = Var\left( \sum_{g=1}^{G} \sum_{i \in A_g} w_i \mu_g \right) + E \left( \sum_{g=1}^{G} \sum_{i \in A_g} (\alpha_i^2 - \alpha_i) \sigma_g^2 \right). \]  

(3.4.5)

The reference distribution in expression (3.4.4) and (3.4.5) is the joint distribution of the superpopulation model, the sampling mechanism, the response mechanism and the imputation mechanism. From (3.4.4), the estimator, \( \hat{Y}_{I,H} \), is an unbiased estimator of the finite population total \( Y_N \) since \( E(\hat{Y}_{I,H} - Y_N|A, A_r, d) = 0 \). Expression (3.4.5) is the variance of the estimator of the finite population total. Under the cell mean model (3.4.2), the conditional variance, \( Var(\hat{Y}_{I,H} - Y_N|A, A_r, d) \), will be minimized when the \( \alpha_i \) are all equal.

Under the cell mean model, if the parameter of interest is only the mean of a variable, then the most efficient estimator of the cell mean is the simple sample cell mean. However, the typical survey situation is that one makes available to different users a completed data set without knowing what kind of parameters the users of the data want to estimate. If the completed data set is produced through preserving the distribution of the observed values, the imputed datasets could be used for general purpose estimation. The fully efficient fractionally imputed estimator (FEFI) is defined as the estimator \( \hat{Y}_{I,H} \) with every respondent in the imputation cell used as a donor for each missing unit in the cell with imputation fractions \( w_{ij}^* \) defined to be proportional to the sampling weights. Since this estimator contains no randomness due to selection of donors, it is called fully efficient.

For variance estimation, Fuller and Kim proposed a consistent replication variance estimation procedure for estimators computed with fractional imputation. The key idea
of constructing an unbiased replication variance estimator for the variance of $\hat{Y}_{I,H}$ in (3.4.5) is to adjust each replicate of the imputation fraction, $w_{ij}^{*(k)}$, to satisfy

$$
\sum_{k=1}^{K} c_k \sum_{i \in A_r} (\alpha_i^{(k)} - \alpha_i)^2 = \sum_{i \in A_r} (\alpha_i^2 - \alpha_i) \ (g = 1, \ldots, G)) \tag{3.4.6}
$$

$$
\sum_{i \in A_r} w_{ij}^{*(k)} = 1 \quad \text{for every } j \in A, \tag{3.4.7}
$$

where $\alpha_i^{(k)} = \sum_{j \in A} w_j^{(k)} w_{ij}^{*(k)}$, $w_j^{(k)}$ is the $k$th replication weight of unit $j$, $w_{ij}^{*(k)}$ is the $k$th replication weight of the imputation fraction $w_{ij}^*$, $K$ is the number of replicates, and $c_k$ is a factor associated with replicate $k$ determined by the replication method. In section 2.4, $\alpha_i^{(k)}$ was denoted $w_{ad,i}^{(k)}$ and $\alpha_i$ was denoted $w_i^*$. Then the replication estimator with adjusted imputation fractional weights is unbiased for the variance of $\hat{Y}_{I,H}$, where the variance is defined by the joint distribution mentioned above. This unbiasedness holds for any survey design for which a replication variance estimator is available. The method of constructing replicates that satisfy (3.4.6) and (3.4.7) is presented in section 4 of their paper. Their theorem 2 demonstrates the unbiasedness of their variance estimator. It is worth noting that only the weights on the imputed values are adjusted in the creation of replicates. In contrast, in Rao and Shao (1992), the adjustments depend on the analysis. In Kim and Fuller (2004), the replicate weights constructed once can be used for estimation of any smooth function of the $y$ variable.

### 3.4.2. Fractional Imputation for the Response Model

Another approach to estimate finite population parameter with missing data is the response model approach. Fuller and Kim (2005a) discussed the use of fractional imputation based on the response model approach without specifying the distribution of the study variable. Response model refers to the assumptions about the probability of obtaining a response from the sample. In this approach, the study variables are
treated as constants and the response probability model is used. A simple but strong assumption for a response model called the response homogeneity groups (RHG) model is commonly used. The RHG model assumes that the population consists of nonoverlapping subgroups, often called cells, and all units in the same cell have the same response probability. Under the RHG model, the respondent data can be viewed as coming from a two-phase sampling procedure where the assumed response model is treated as the second phase sampling mechanism. In this setting, a reasonable estimator of the total is a reweighted expansion estimator (REE),

\[
\hat{Y}_{REE} = \sum_{g=1}^{G} \left( \sum_{i \in A_g} w_i \right) \frac{\sum_{i \in A_{rg}} w_i y_i}{\sum_{i \in A_{rg}} w_i}. \tag{3.4.8}
\]

This is formula (7) of Fuller and Kim (2005a). Fuller and Kim (2005a) labeled the REE estimator as a fully efficient estimator. The label is appropriate because it contains no variability due to random selection of donors. As with the construction of the fully efficient fractionally imputed (FEFI) estimator in Kim and Fuller (2004), the fractionally imputed estimator is constructed to be algebraically equivalent to REE (3.4.8) through the use of every respondent in an imputation cell as a donor for every nonrespondent in the same cell and defining the imputation fractions to be proportional to the sampling weights. Since their FEFI estimator,

\[
\hat{Y}_{I,H} = \sum_{g=1}^{G} \sum_{j \in A_g} \sum_{i \in A_{rg}} w_j w_{ij}^* y_i, \tag{3.4.9}
\]

is algebraically equal to the REE estimator, the theoretical results for REE can be used for properties of the FEFI estimator.

Kim, Navarro and Fuller (2006) showed some asymptotic properties of the REE under a sequence of populations and samples increasing in size. Let the populations be indexed by \( t = 1, 2, \ldots \) such that the populations are nested: \( U_1 \subset U_2 \subset \ldots \). Then the populations sizes are increasing: \( N_1 < N_2 < \ldots \). It is assumed that the sample sizes are increasing (\( n_1 < n_2 < \ldots \)), but the samples are not necessarily nested. Assume
that the population is composed of $G$ mutually exclusive and exhaustive cells. Assume that the variance of a full sample estimator of the total in population $t$ is $O(n_t^{-1}N_t)$.

Under regularity conditions and the assumption that the responses are independent, the procedures used by Kim, Navarro and Fuller (2006) in the proof of their Theorem 2.1 can be used to show that the REE (3.4.8) satisfies

$$
\hat{Y}_{REE} = \tilde{Y}_{REE} + o_p(n_t^{-1}N_t),
$$

where

$$
\tilde{Y}_{REE} = \hat{Y}_{nt} + \sum_{g=1}^{G} \sum_{i \in A_g} w_i (p_g^{-1}R_i - 1)e_i
$$

(3.4.11)

where $e_i = y_i - \bar{Y}_g$, $\bar{Y}_g = \sum_{i \in U_g} y_i$ is the population mean in cell $g$, and $p_g = Pr(R_i = 1|i \in A_g)$ is the response probability in cell $g$. The reference distribution in (3.4.10) is the joint distribution of the sampling mechanism and the response mechanism, conditional on the realized values of $(y_i, R_i)$ in the sample. Expression (3.4.10) states that $\hat{Y}_{REE}$ is asymptotically equivalent to $\tilde{Y}_{REE}$.

The variance of $\tilde{Y}_{REE}$ is

$$
V(\tilde{Y}_{REE}) = V(\hat{Y}_n) + E \left( \sum_{g=1}^{G} \sum_{i \in A_g} p_g^{-1}(1 - p_g)w_i^2 e_i^2 \right).
$$

(3.4.12)

This is formula (10) of Fuller and Kim (2005a). The variances on (3.4.12) are with respect to the joint distribution defined by the sampling mechanism and the response mechanism. Correspondingly the expectation is with respect to the joint distribution. The first term of the right-hand side is the ordinary sampling variance of the complete sample. The second component is the variance due to the response mechanism. Expression (3.4.12) shows that the key factors determining the conditional variance due to the response mechanism are the response probabilities and the weighted residuals. The closer $p_g$ is to 1, the closer the conditional variance, conditional on the observed sample indices, is
to 0. The residuals also affect the variance. The sum of the squared weighted residuals will be small if the predicted values are close to the actual values.

Fuller and Kim also provided a computationally efficient replication variance estimator of FEFI estimator. Each replicate of the imputation fraction, \( w^{*^{(k)}}_{ij} \), is to be adjusted in such a way that each replicate for the FEFI estimator is algebraically equivalent to each replicate of the REE (3.4.8). Formally, each replicate, \( w^{*^{(k)}}_{ij} \) of the imputation fraction is created by satisfying

\[
\hat{Y}^{(k)}_{I,H} = \sum_{g=1}^{G} \sum_{j \in A_g} \sum_{i \in A_{rg}} w^{(k)}_{ij} w^{*^{(k)}}_{ij} y_i,
\]

\[
\equiv \sum_{g=1}^{G} \left( \sum_{i \in A_g} w^{(k)}_{i} \right) \frac{\sum_{i \in A_{rg}} w^{(k)}_{i} y_i}{\sum_{i \in A_{rg}} w^{(k)}_{i}} = \hat{Y}^{(k)}_{REE}.
\]

Using the replicates (3.4.13), the replicate variance estimator can be written as

\[
\hat{V}_{JK}(\hat{Y}_{I,H}) = \sum_{k=1}^{L} c_k (\hat{Y}^{(k)}_{I,H} - \hat{Y}_{I,H})^2.
\]

Fuller and Kim (2005a) showed that \( \hat{V}_{JK}(\hat{Y}_{I,H}) \) satisfies

\[
\hat{V}_{JK}(\hat{Y}_{I,H}) = V(\hat{Y}_{REE}) - \sum_{g=1}^{G} \sum_{i \in U_g} p^{-1}_g (1 - p_g) \epsilon^2_i + o_p(n_t^{-1}N_t),
\]

where the distribution is with respect to the sampling and response mechanisms.

The replication estimator (3.4.14) is consistent for the variance of the imputed estimator when the finite population correction is ignorable. If not, an estimator of the second term on the right side of (3.4.15) should be added to the replication estimator. This is expressed in formula (17) of Fuller and Kim (2005a). Fully efficient fractional imputation may not be used so commonly because of the size of the resulting data set. Fuller and Kim (2005a) also suggested a procedure with a small number of donors per recipient that is fully efficient for the total with modification of the imputation fraction \( w^{*}_{ij} \) using the regression weighting technique.
Until now, we described fractional hot deck imputation through two different approaches. These approaches are ideal for simple situations in which there are no additional variables in the imputation cell. In other words, the information contained in the cell are only used to construct imputation cells. A weighting system which makes better use of the auxiliary information can be developed for any case where there are additional variables present within an imputation cell. For incorporating information contained in all auxiliary variables into the imputed estimator, Kim (2007) suggests the regression weighting technique. More detail of Kim’s paper will be described in next section.

3.4.3. Fractional Regression Hot Deck Imputation

Kim and Fuller (2004) described the fractional hot deck imputation method and introduced a preliminary weighting system for this method. When the sampling weights in a cell are the same, all imputed values have the same weight $M^{-1}$, where $M$ is the number of donors. In this weighting system, the imputation fraction $w_{ij}^*$ does not make use of the information contained in an auxiliary variable associated with the units within an imputation cell. Kim (2007) suggested a weighting technique which uses all auxiliary variables in an imputation cell with missing observations based on a population model approach.

In many practical cases, the missing value is replaced by a value predicted from an assumed model plus a residual term. The residual term is added to preserve the marginal variability of the original data after imputation. The completed data set made by the regression imputation tends to have less variability than a set of truly observed values $y_i$ because the regression fit that gives $\hat{y}_i = x_i'\hat{\beta}$ is to some degree a result of data smoothing. Adding a residual will alleviate this problem. As a result, the completed data by adding a randomly selected residual come closer to displaying the natural amount of variation. When the idea of fractional imputation is applied with this imputation
scheme, the imputed estimator of total can be constructed as

\[ \hat{Y}_{I, re} = \sum_{i \in A_r} w_i y_i + \sum_{j \in A_m} \sum_{i \in A_r} w_j w_{ij}^{*} (\hat{y}_j + \hat{e}_i) \]  

(3.4.16)

where \( \hat{y}_j \) is the predicted value of \( y_j \) and \( \hat{e}_i = y_i - \hat{y}_i \) is an imputed residual. Under model (3.3.1), the weighted mean of the imputed values can be written

\[ \sum_{i \in A_r} w_j w_{ij}^{*} (\hat{y}_j + \hat{e}_i) = \sum_{i \in A_r} w_{ij}^{*} y_i \]  

(3.4.17)

if the imputation fraction \( w_{ij}^{*} \) satisfies

\[ \sum_{i \in A_r} w_{ij}^{*} (1, x_i') = (1, x_j') \text{ for } j \in A_m. \]  

(3.4.18)

The right-hand in (3.4.17) is an expression for the weighted mean of hot-deck imputed values. The solution that satisfies (3.4.18) was given as the regression weighting equation in section 2.7 of this document, which is a reproduction of Kim’s (2007) formula (2.6).

Under (3.4.18), the regression imputation estimator \( \hat{Y}_{I, re} \) in (3.4.16) has the form of hot deck fractional imputation (3.4.3). Therefore, Kim’s method can be viewed as fractional hot deck imputation with a weighting technique which uses auxiliary variables. The weighing technique is made to preserve the correlation structure of the important variables.

Under model (3.3.1) and an ignorable response mechanism, the conditional expectation of the regression imputation estimator \( \hat{Y}_t \) of \( Y_N \) is

\[ E(\hat{Y}_{I, re} - \hat{Y}_n | X, R, A) = E \left( \sum_{j \in A_m} w_j \left[ \sum_{i \in A_r} w_{ij}^{*} x_i' \beta_0 - y_j \right] | X, R, A \right) \]  

(3.4.19)

\[ = \sum_{j \in A_m} w_j \left( \sum_{i \in A_r} w_{ij}^{*} x_i' \beta_0 - x_j' \beta_0 \right), \]

which is equal to zero by (3.4.18), where \( X = \{(i, x_i) : i \in A\}, R = \{(i, R_i) : i \in A\} \). The reference distribution above is the joint distribution of the superpopulation model (3.3.1), the sampling mechanism, and the response mechanism. The imputed estimator (3.4.16)
with adjusted imputation fractions is unbiased since the complete sample estimator is unbiased under the sampling mechanism.

The suggested replication variance estimator in Kim (2007) is closely related to that of Kim and Fuller (2004) that adjusts the fractional replicate weights to produce a consistent estimator of the variance. Each replicate of the imputation fraction, $w_{ij}^{* (k)}$, is adjusted to satisfy

$$\sum_{k=1}^{L} c_k \sum_{i \in A_r} (\alpha_i^{(k)} - \alpha_i)^2 = \sum_{i \in A_r} \alpha_i^2$$

(3.4.20)

and

$$\sum_{i \in A_r} w_{ij}^{* (k)}(1, x_i') = (1, x_j') \text{ for } j \in A.$$ (3.4.21)

The procedure used by Kim (2007) is very similar to the previously described procedures for adjusting weights. Kim (2007) considers fractional weights in replicate $k$ in the following class:

$$w_{ij}^{* (k)} = \begin{cases} 
   w_{ij}^* - (1 - w_{ij0}^*)b_k + h_{ik,j}b_k & \text{if } d_{kj} = 1, d_{ij} = 1 \text{ and } k = i \\
   w_{ij}^* + w_{ij0}^*b_k + h_{ik,j}b_k & \text{if } d_{kj} = 1, d_{ij} = 1 \text{ and } k \neq i \\
   w_{ij}^* & \text{otherwise,}
\end{cases}$$ (3.4.22)

where

$$h_{ik,j} = \alpha_{ij}(x_i - \bar{x}_{Ij})S_{xx,j}^{-1}(x_k - \bar{x}_{Ij}),$$

$$S_{xx,j} = \sum_{i \in A_{rj}} \alpha_{ij}(x_i - \bar{x}_{Ij})^2,$$

$$\bar{x}_{Ij} = \sum_{i \in A_r} \alpha_{ij}x_i,$$

and $b_k$ is to be determined. The value of $b_k$ is determined by solving quadratic equation (3.8) of Kim (2007). As with the FHDI method of Kim and Fuller (2004), the FRHDI method of Kim (2007) produces an unbiased variance estimation method.
4. Fractional Deterministic Imputation (FDI)

Imputation is the substitution of some plausible values for the values of items that are missing. In general, the successful imputation is conditional on the use of observed auxiliary information. This will reduce the nonresponse bias and the variance. When all missing values have been imputed, one can then analyze the imputed data sets using standard techniques for complete data. Different users can have consistent results. If each user, however, takes a different approach to imputation, results can vary. The imputed values can be obtained from a statistical prediction rule or observed values having similar auxiliary variables. The imputed values can be classified into deterministic (the exactly same imputed values are constructed when repeating the imputation procedure on the same data set) or stochastic (random imputed values are produced). The imputation method using a deterministic rule is called “Deterministic Imputation”, whereas “Stochastic Imputation” denotes imputing the value using a stochastic rule. For example, regression imputation can be viewed as a Deterministic Imputation method, whereas regression imputation plus a random residual can be viewed as a Stochastic Imputation method.

In this paper, we focus on Deterministic Imputation (DI) that is commonly used in many surveys. Under a common DI scheme, a missing value is replaced by the predicted value obtained by fitting an assumed model using respondent values in the sample. The imputed estimator based on the completed data sets constructed by the DI method is often unbiased and provides efficient estimates for estimating functions of marginal population totals when the assumed model holds. However, there are some potential disadvantages of the DI method. It distorts the distribution of the variable that has to be imputed and thus the variance is underestimated. In addition, the correlation between the imputed data sets is artificially inflated.

To preserve the variances and correlation in the imputed data, stochastic imputation
methods are suggested. A common way to conduct stochastic imputation (STI) is to add a residual to the predicted value of the DI. The completed data constructed by STI comes closer than DI to displaying the natural amount of variation in the observed data. The disadvantage of this method is that it may be undesirable to estimate marginal totals due to the added imputation variance.

The main purpose of this paper is to propose a new imputation procedure for providing an easy-to-use data set to the end user. The resulting estimator of a population total based on the completed data set obtained by the proposed method is asymptotically equivalent to the DI estimator. The proposed imputation method provides approximately unbiased and efficient estimates for the variance and correlation, a property that DI estimators do not achieve. It retains many of the desirable properties of the deterministic imputation scheme for many parameters of interest. For more convenience, the construction of general purpose replicates for variance estimation is also studied. It is well known that considering imputed values as if they were true values may lead to serious underestimation of the variance of the imputed estimator. To avoid this underestimation, the end user could use more advanced techniques, such as the adjusted jackknife method of Rao and Shao (1992), variance estimation via a decomposition method (Deville and Särndal (1994)), or the linearization and reverse approach of Kim and Rao (2009). Some end users, however, will not be able to use such advanced techniques. It is possible with our procedure for the end user to obtain reasonable variance estimates for the variance of the imputed estimator using standard replication methods based on the single set of replicate weights.

4.1. Introduction

Let \( U = \{1, 2, \ldots, N\} \) be a set of indices of a finite population \( F_N \) of \( N \) elements with \( N \) known. Associated with the \( i^{\text{th}} \) unit of \( U \) is \( y_i \), a variable of interest, and \( x_i \), an auxiliary variable (or vector of variables). Let \( A \) denote the set of indices of the elements
in a sample of size \( n \) selected by the chosen sampling mechanism. For \( i \in A \), assume that some \( y_i \) are missing, but that every \( x_i \) is observed. Responses \( y_i \) are obtained from the selected sample according to the response mechanism. Let \( A_r \) be the set of respondents and \( A_m \) the set of nonrespondents. In this notation, \( A = A_r \cup A_m \) and \( A_r \cap A_m = \emptyset \).

Under nonresponse, we define the response indicator variable for \( y_i \) by

\[
R_i = \begin{cases} 
1 & \text{if } y_i \text{ responds,} \\
0 & \text{if } y_i \text{ does not respond.}
\end{cases}
\]

Assume that the response mechanism is missing at random in the sense that the response probability does not depend on the study variable being imputed but may depend on auxiliary variables used for imputation.

Let the population quantity of interest be \( \theta_N = \theta(y_1, y_2, \ldots, y_N) \). Let \( \hat{\theta}_1n \) be a design linear estimator of \( \theta_1N = \sum_{i=1}^{N} y_i \) based on the full sample,

\[
\hat{\theta}_1n = \sum_{i \in A} w_i y_i,
\]

(4.1.1)

where \( w_i = (Pr(i \in A))^{-1} \) is a sampling weight for unit \( i \) that depends on the sampling mechanism. In addition to the population total, suppose that we are also interested in the population correlation between \( x \) and \( y \) and marginal variance of \( y \). For simplicity, we assume that other parameters of interest are \( \theta_2N = \sum_{i=1}^{N} x_i y_i \) and \( \theta_3N = \sum_{i=1}^{N} y_i^2 \).

The complete sample estimators of \( \theta_2N \) and \( \theta_3N \) are

\[
\hat{\theta}_2n = \sum_{i \in A} w_i x_i y_i,
\]

(4.1.2)

and

\[
\hat{\theta}_3n = \sum_{i \in A} w_i y_i^2.
\]

(4.1.3)

We assume that \( Pr(i \in A) > 0 \) for all \( i \), so that

\[
E(\hat{\theta}_n|F_N) = \theta_N,
\]

(4.1.4)
where $E(\cdot | F_N)$ is the design expectation and denotes the average over all samples possible under the design for the given finite population $F_N$. The estimators $\hat{\theta}_1n$, $\hat{\theta}_2n$ and $\hat{\theta}_3n$ are the traditional Horvitz-Thompson estimators.

In the presence of missing data, an estimator for $\theta_1N$ is,

$$\hat{\theta}_r1 = \sum_{i \in A} w_i \{ R_i y_i + (1 - R_i) E(y_i|x_i) \}. \quad (4.1.5)$$

where the expectation is with respect to the conditional distribution of $y$ given $x$. Some assumption is needed to compute $E(y_i|x_i)$ in practice. Under MAR, $\hat{\theta}_r1$ should be unbiased for $\theta_1N$. Estimator of $\theta_2N$ and $\theta_3N$ can be defined similarly. Let

$$\hat{\theta}_r2 = \sum_{i \in A} w_i \{ R_i x_i y_i + (1 - R_i) E(x_i y_i|x_i) \} \quad (4.1.6)$$

$$= \sum_{i \in A} w_i \{ R_i x_i y_i + (1 - R_i) x_i E(y_i|x_i) \}.$$ 

and

$$\hat{\theta}_r3 = \sum_{i \in A} w_i \{ R_i y_i^2 + (1 - R_i) E(y_i^2|x_i) \}. \quad (4.1.7)$$

In general, an analytical method is used to compute the conditional expectation in the above estimators. In addition, numerical methods such as a Monte Carlo method also are used to approximate the conditional expectation. The Monte Carlo approximation method can be viewed as an imputation method. Recently, Kim and Fuller (2009) proposed the parametric fractional imputation using the idea of Monte Carlo approximation. For example, by using a Monte Carlo method, the conditional expectation can be approximated by

$$E(y|x_i) \approx \frac{1}{T} \sum_{t=1}^{T} y_{i}^{*}(t), \quad (4.1.8)$$

where $y_{i}^{*}(t)$ are generated from the conditional distribution $f(y|x_i)$. For generating $y_{i}^{*}(t)$, the model of the conditional distribution is needed and can be parametric, semi-parametric, or non-parametric. In the parametric approach, $f(y|x_i, \beta)$ is assumed for
the unknown conditional distribution and \( y^{*t} \) are generated from the conditional distribution \( f(y|x_i, \hat{\beta}) \) with an estimated value \( \hat{\beta} \). This is the main approach of parametric fractional imputation proposed by Kim and Fuller (2009). Parametric imputation is efficient but is not robust against the misspecification of the assumed model. In practice, the assumption of a conditional distribution may not be possible. Fortunately, the assumption of some moments may be more easily constructed based on prior information. When only some moment assumptions are made, however, the imputed values cannot be generated from the conditional distribution since we do not have the exact form of \( f(y|x) \). In this case, a semi-parametric imputation method can be considered for missing data analysis.

In this paper, we consider a semi-parametric imputation method using empirical likelihood. The method is constructed so that it shares the desirable properties of the DI method. In section 4.2, we discuss deterministic imputation and an assumption of a semi-parametric model for the distribution of \( y \) given \( x \). In section 4.3, we propose a new method, fractional deterministic imputation (FDI). The method should retain the consistency of deterministic imputation and enable variance estimation. In section 4.4, we address the issue of constructing imputation fractions. In section 4.5, we discuss some theory for the FDI method. In section 4.6, we consider variance estimation for FDI. In section 4.7, we present a simulation study.

4.2. Deterministic Imputation: DI

In this paper, we assume some moment conditions:

\[
E(y_i|x_i) = g(x_i, \beta_0),
\]

\[
V(y_i|x_i) = \sigma^2 q(x_i, \beta_0),
\]

and

\[
cov(y_i, y_j|x_i) = 0 \text{ if } i \neq j
\]
for some $p$-dimensional vector $\beta_0$, where $g(x_i, \beta)$ and $q(x_i, \beta)$ are known functions of $x_i$ for given a $\beta$. Under a semi-parametric model, Deterministic method can be considered. We assume that the response mechanism is missing at random. The parameter estimate $\hat{\beta}$ is taken to be the solution of

$$
\hat{U}(\beta) \equiv \sum_{i \in A_r} w_i \{y_i - g(x_i, \beta)\} h_i(\beta) = 0,
$$

(4.2.2)

where $h_i(\beta) = \dot{g}(x_i, \beta)/q(x_i, \beta)$ and $\dot{g}(x_i, \beta) = \partial g(x_i, \beta)/\partial \beta$. This choice is suggested by Kim and Rao (2009) based on the quasi-likelihood equations for generalized linear models. We assume that the solution $\hat{\beta}$ is unique.

In deterministic imputation (DI), the conditional expectation $E(y_i|x_i)$ is replaced by $g(x_i, \hat{\beta})$. Under DI method, an imputed value for a missing $y_i$ is

$$
y_i^* = g(x_i, \hat{\beta}).
$$

(4.2.3)

The imputed estimators for the population parameters based on the completed data constructed by (4.2.3) can be written as

$$
\hat{\theta}_{DI,1n} = \sum_{i \in A_r} w_i y_i + \sum_{j \in A_m} w_j \hat{y}_j,
$$

(4.2.4)

$$
\hat{\theta}_{DI,2n} = \sum_{i \in A_r} w_i x_i y_i + \sum_{j \in A_m} w_j x_j \hat{y}_j,
$$

and

$$
\hat{\theta}_{DI,3n} = \sum_{i \in A_r} w_i y_i^2 + \sum_{j \in A_m} w_j x_j \hat{y}_j^2.
$$

Under the assumed model, the DI estimator for the population total is conditionally unbiased and quite efficient. However, the completed data set made by (4.2.3) tends to have less variability than a set of truly observed values $y_i$ because the fit that gives $\hat{y}_i$ is to some degree a result of data smoothing. As a result, the population correlation coefficient and the variance of the study variable are over-estimated and under-estimated, respectively. In addition, estimation based on the completed data set for population
quantiles relatively far from the mean of the observed study variable is subject to under-estimation and relatively close to the observed mean is subject to over-estimation. To avoid those problems, adding residual noise is suggested in the literature. The completed data by adding randomly selected residuals come closer to displaying the natural amount of variation. However, the efficiency of estimators will be reduced. Since end users may consider completed data sets as a true observed sample and analyze the provided data set to estimate other parameters, those disadvantages lead end users to invalid inference.

As pointed by Kim and Rao (2009), the parameter $\beta$ can be viewed as a nuisance parameter since our main interest of parameter is the population parameter not $\beta$ itself. Therefore, the sample variability of the estimated nuisance parameter can be explained by Taylor expansion method. To further discuss the asymptotic properties of the DI estimators, we assume that the finite population $U$ is a member of a sequence of finite populations indexed by $v$. All limiting processes in this paper are understood to be as $v \to \infty$. As $v \to \infty$, the population size $N$ and the sample size $n$ increase to infinity. Further assume that

[C1] \[
\max_{i \in U} \frac{nw_i}{N} < K_0
\]
for some $K_0 > 0$.

[C2] The sampling design is such that, for any $z_i$ with bounded $(2 + \delta)$ moments,

\[
n \var\left( N^{-1} \sum_{i \in A} w_i z_i | F_N \right) < K_1
\]

for some $K_1 > 0$, where $F_N = \{z_1, \ldots, z_N\}$.

[C3] For each $i$, $g(x_i, \beta)$ and $h(x_i, \beta)$ are continuous functions of $\beta$ in a compact set $B$ containing $\beta_0$ as an interior point.
For each \( i \), \( g(x_i, \beta) \) is differentiable with continuous partial derivative \( \dot{g}(x_i, \beta) \) in a compact set containing \( \beta_0 \) and \( \sum_{i=1}^{N} R_i \dot{g}(x_i, \beta_0)h(x_i, \beta_0)^T \) is nonsingular.

The equation (4.2.2) has a unique solution in a bounded subset of the parameter space which contains \( \beta_0 \), the value of \( \beta \) used to generate the infinite population

Conditions [C1] and [C2] are the standard condition for the sequence of finite populations and samples. Conditions [C3] and [C4] are sufficient conditions for showing the uniform convergence of some functions of the estimated nuisance parameter and the consistency of \( \hat{\beta} \). Condition [C5] ensures that there is an unique estimate of \( \beta \) in the parameter space and that \( \beta \) can be consistently estimated by solving the estimating equation (4.2.2).

The DI method provides a consistent estimate for the marginal total \( \theta_{1N} \) and the covariance of \( y \) and \( x \) since \( \hat{\theta}_{DI,1n}^I \) and \( \hat{\theta}_{DI,2n}^I \) are consistent estimators for \( \theta_{1N} \) and \( \theta_{2N} \), respectively. It should be possible to show that \( \hat{\theta}_{DI,3n}^I \) is not a consistent estimator of \( \theta_{3N} \). Indeed, the estimator \( \hat{\theta}_{DI,3n}^I \) should have an asymptotic bias. The reason that there is this bias and a lack of consistency is that the estimator uses the squared predicted value of \( y \) as an estimator of the square of \( y \) instead of an estimator of the actual expected value of \( y \) conditional on \( x \). As the sample size grows large, the bias equal to the

\[-E \left( \sum_{i=1}^{N} (1 - R_i)\sigma^2 q(x_i, \beta_0) \right)\]

remains. In addition to assumptions [C1]-[C5], one would assume that the super-population model has \((2 + \delta)\) moments of \((x_i, y_i, g(x_i, \beta_0), \dot{g}(x_i, \beta_0), h(x_i, \beta_0), \dot{h}(x_i, \beta_0))\) for some \( \delta > 0 \), where \( h(x_i, \beta) = \dot{g}(x_i, \beta)/q(x_i, \beta) \), \( \dot{g}(x_i, \beta) = \partial g(x_i, \beta)/\partial \beta \) and \( \dot{h}(x_i, \beta) = \partial h(x_i, \beta)/\partial \beta \) and \( \beta_0 \) satisfies the model (4.2.1). Also, one assumes that the sampling mechanism and the response mechanism are ignorable under the model.

Consequently, \( \hat{\theta}_{DI,3n}^I \) in (4.2.4) is not unbiased for the population parameter \( \theta_{3N} \). This result implies that the completed data set using the DI method tends to have less variability than a set of observed values. In other words, the deterministic imputation
method using (4.2.3) cannot preserve the distribution of the observed variables. As a result, it is not an appropriate imputation method for other characteristics of the distribution of a study variable.

4.3. Proposed Method: FDI

Suppose that the conditional distribution, $f(y|x_j)$, has support on \{y_1^*, y_2^*, \ldots, y_T^*\} for some $y_1^*, y_2^*, \ldots, y_T^*$. Define

$$ p_{ij} = \frac{f(y_i^*|x_j)}{\sum_{i=1}^T f(y_i^*|x_j)} \quad (4.3.1) $$

be the unknown point mass assigned to $y_i^*$ for $f(y|x_j)$. By the definition of $p_{ij}$, the following equation can be considered:

$$ \sum_{i=1}^T p_{ij}(1, y_i^*, y_i^{*2}) = (1, E(y|x_j), E(y^2|x_j)) \quad (4.3.2) $$

Note that $p_{ij}$ takes the role of fractional weights.

Consider the semi-parametric model (4.2.1). We cannot compute $p_{ij}$ defined in (4.3.1) since the exact form of $f(y|x)$ is not known. Under the assumed model, $E(y|x_j) = g(x_j, \beta_0)$ and $E(y^2|x_j) = g(x_j, \beta_0)^2 + \sigma^2 q(x_j, \beta_0)$. Assume that $\hat{\beta}$ and $\hat{\sigma}$, consistent estimators of $\beta_0$ and $\sigma$, are available. As with the idea of the DI method, our goal is to find weights $p_{ij}$ to satisfy

$$ \sum_{i=1}^T p_{ij}(1, y_i^*, y_i^{*2}) = (1, g(x_j, \hat{\beta}), g(x_j, \hat{\beta})^2 + \hat{\sigma}^2 q(x_j, \hat{\beta})) \quad (4.3.3) $$

One possible choice for $\hat{\beta}$ is the solution to (4.2.2). A choice for $\hat{\sigma}^2$ is

$$ \hat{\sigma}^2 = \left( \sum_{i \in A} w_i \right)^{-1} \sum_{i \in A} w_i \frac{(y_i - g(x_i, \hat{\beta}))^2}{q(x_i, \hat{\beta})}. \quad (4.3.4) $$

To find $p_{ij}$ satisfying (4.3.4) is a typical example of calibration estimation using the moment conditions in sample surveys.
The empirical likelihood method can be used to compute the fractional weights. If the entire support $y_1^*, y_2^*, \ldots, y_T^*$ on $f(y|x_j)$ is known, we consider the empirical likelihood to be $\prod_{i=1}^T p_{ij}$. The corresponding log likelihood function:

$$l(p) = \sum_{i=1}^T \log(p_{ij}). \quad (4.3.5)$$

We produce fractional weights $p_{ij}$ by maximizing $l(p)$ subject to constraints. The procedure can be viewed as an example of calibration estimation and approached using Lagrange multipliers. In practice, we do not know the entire support. To overcome the difficulty of obtaining $y^*$ for the entire support, we construct the donor set $A_{D_j}$ of size $M$ where each element is assumed to be a support of the conditional distribution.

Through viewing the log-likelihood function in (4.3.5) as a finite population total, a reasonable estimate of $l(p)$ is

$$\hat{l}(p) = \sum_{i \in A_{D_j}} w_{ij0}^* \log(p_{ij}), \quad (4.3.6)$$

where $w_{ij0}^*$ are the initial weights. This approach is motivated by pseudo empirical likelihood (Chen and Sitter, (1999)). Then the fractional weights can be obtained from maximizing (4.3.6) subject to

$$\sum_{i \in A_{D_j}} p_{ij}(1, y_i, y_i^2) = \left(1, g(x_j, \hat{\beta}), g(x_j, \hat{\beta})^2 + \hat{\sigma}^2 q(x_j, \hat{\beta})\right). \quad (4.3.7)$$

Using the Lagrange multiplier method, for $i \in A_{D_j}$ and $j \in A_m$, the estimator of $p_{ij}$ is

$$\hat{p}_{ij} = \frac{w_{ij0}^*}{1 + \lambda^T u_i}, \quad (4.3.8)$$

where

$$u_i^T = \left(y_i - g(x_j, \hat{\beta}), y_i^2 - [g(x_j, \hat{\beta})^2 + \hat{\sigma}^2 q(x_j, \hat{\beta})]\right)$$

and the Lagrange multiplier, $\lambda$, is the solution to

$$\sum_{i \in A_{D_j}} w_{ij0}^* \frac{u_i}{1 + \lambda^T u_i} = 0.$$
More discussion of the Lagrange multiplier method is presented in the next section.

Once the fractional weights are computed in this manner, the imputed estimators of \( \theta_1N \), \( \theta_2N \) and \( \theta_3N \) can be constructed by

\[
\hat{\theta}_{FDI,1n} = \sum_{i \in A_r} w_i y_i + \sum_{j \in A_m} \sum_{i \in AD_j} w_j \hat{p}_{i|j} y_i, \tag{4.3.9}
\]

\[
\hat{\theta}_{FDI,2n} = \sum_{i \in A_r} w_i x_i y_i + \sum_{j \in A_m} \sum_{i \in AD_j} w_j \hat{p}_{i|j} x_j y_i,
\]

and

\[
\hat{\theta}_{FDI,3n} = \sum_{i \in A_r} w_i y_i^2 + \sum_{j \in A_m} \sum_{i \in AD_j} w_j \hat{p}_{i|j} y_i^2.
\]

To summarize, for conducting our imputation procedure, we consider the following steps for creating completed data sets.

**STEP 1.** Compute \( \hat{\beta} \) and \( \hat{\sigma}^2(\hat{\beta}) \).

**STEP 2.** Select \( M \) donors for each missing unit

**STEP 3.** Compute the imputation fractions satisfying condition (4.3.7) using \( \hat{\beta} \) and \( \hat{\sigma}^2(\hat{\beta}) \) computed from **STEP 1**.

In step 2, various methods of selecting donors can be used. These include systematic sampling from the available donors sorted in some manner, simple random sampling without replacement from the observed cases, or selection of donors using some nearest neighbor method. Selecting donors is an important step in the procedure. Careful selection should be able to avoid extreme fractional weights.

Each donor set for each \( j \in A_m \) may be constructed by using observed values in the sample or pseudo values. A difficulty may arise when (4.3.7) does not have a proper solution where the convex hull of \( \{y_i, y_i^2, i \in AD_j\} \) does not contain the target values, \((g(x_j, \hat{\beta}), g(x_j, \hat{\beta})^2 + \hat{\sigma}^2 q(x_j, \hat{\beta}))\). To guarantee the existence of a solution, each donor
set $A_D$ should be constructed so that it has in its convex hull $(g(x_j, \hat{\beta}), g(x_j, \hat{\beta})^2 + \hat{\sigma}^2 q(x_j, \hat{\beta})))$. In some situations, a donor set constructed using observed $y \in A_r$ cannot satisfy the condition of convex hull. In such a situation, the solution could have negative weights included in the single weight set provided to the end user. Since some end users may not accept negative weights, a method of avoiding negative weights should be considered.

This problem can be avoided by using pseudo support values. We do not have the exact form of $f(y|x)$ but some moment assumptions. Using the assumptions of moments, we may define a reasonable range of support. For missing unit $j$, define $a = \hat{y}_j - z\hat{\sigma}^2 q(x_j, \hat{\beta})$ and $b = \hat{y}_j + z\hat{\sigma}^2 q(x_j, \hat{\beta})$, where $z$ is defined by selecting a reasonable bound for support. Then define the interval $[a, b]$ be subdivided into $M_2$ subintervals, $\{[a_{t-1}, a_t]\}_{t=1}^{M_2}$ of equal width $h = (b - a)/M_2$ where $a_0 = a, a_{M_2} = b$. Form equally spaced nodes $s_t = a + (t - 1/2)h$ for $t = 1, 2, \ldots, M_2$. A sample of size $M$ is selected by systematic sampling from the set $\{s_t : t = 1, 2, \ldots, M_2\}$. This sample is considered as a donor set for missing unit $j$. Since the constructed donor sets contain the target values, non-negative weights are guaranteed.

4.4. Constructing imputation fractions

In order to use the proposed fractional imputation method, one must construct weights satisfying (4.3.7) using a calibration technique. A calibration technique chooses the adjusted weights that minimize a function of the weights subject to some restrictions. Various calibration techniques are available in the literature. Deville and Särndal (1992) proposed a general calibration method in survey sampling. For our case, it is important to avoid extreme weights because applying these weights to make estimates for proportions may produce unrealistic estimates. In addition, the end users may not accept negative weights so that the method of getting positive fractional weights satisfying equation (4.3.7) should be studied for a general purpose imputation procedure.
In this section, we consider the method of constructing fractional weights to avoid the chance of extreme weights, including negative weights, using a calibration technique to obtain range restricted weights.

In our situation, the fractional weights can be obtained by solving

\[
\maximize \sum_{i \in A} w^*_ij0 \log(p_{ij}) \tag{4.4.1}
\]

subject to

\[
\sum_{i \in A_{Dj}} p_{ij}(1, y_i, y_i^2) = \left(1, g(x_j, \hat{\beta}), g(x_j, \hat{\beta})^2 + \hat{\sigma}^2 q(x_j, \hat{\beta})\right), \tag{4.4.2}
\]

where \(w^*_ij0\) is an initial weight satisfying \(\sum_{i \in A_{Dj}} w^*_ij0 = 1\). In the typical situation with \(M\) donors for missing \(j\), the \(w^*_ij0\) are all equal to \(M^{-1}\). This problem can be solved by using a calibration technique to minimize a function of the distance between an initial weight \(w^*_ij0\) and \(p_{ij}\) subject to the restriction (4.4.2). The following distance measure,

\[
d(p_{ij}, w^*_ij0) = \sum_{i \in A_{Dj}} w^*_ij0 \log \left( p_{ij} \right),
\]

can be defined to construct nonnegative fractional weights. This choice is motivated by the pseudo empirical maximum likelihood method for unequal probability designs (Chen and Sitter, (1999)). Then the objective function to be minimized is

\[
Q(p) = d(p_{ij}, w^*_ij0) - \lambda_1 \left\{ \sum_{i \in A_{Dj}} p_{ij}y_i - g(x_j, \hat{\beta}) \right\} - \lambda_2 \left\{ \sum_{i \in A_{Dj}} p_{ij}y_i^2 - [g(x_j, \hat{\beta})]^2 + \hat{\sigma}^2 q(x_j, \hat{\beta}) \right\} - \lambda_0 \left\{ \sum_{i \in A_{Dj}} p_{ij} - 1 \right\}, \tag{4.4.3}
\]

where \(\lambda\) are Lagrange multipliers.

The equation defining the weights by setting the partial derivative of \(Q\) with respect to \(p_{ij}\) equal to zero is

\[
\frac{\partial Q}{\partial p_{ij}} = \frac{w^*_ij0}{p_{ij}} - \lambda_1 y_i - \lambda_2 y_i^2 - \lambda_0 = 0. \tag{4.4.4}
\]
Then using $\sum_{i \in A_D_j} p_{ij} \left( \frac{\partial Q}{\partial p_{ij}} \right) = 0$, $\sum_{i \in A_D_j} w_{ij0}^* = 1$, and the constraints (4.4.2), we have

$$\lambda_0 = 1 - \lambda_1 t_{1j} - \lambda_2 t_{2j}$$  \quad (4.4.5)$$

and

$$p_{ij} = \frac{w_{ij0}^*}{1 + \lambda_1(y_i - t_{1j}) + \lambda_2(y_i^2 - t_{2j})}, \quad (4.4.6)$$

where $t_{1j} = g(x_j, \hat{\beta})$, $t_{2j} = g(x_j, \hat{\beta})^2 + \hat{\sigma}^2 q(x_j, \hat{\beta})$, and the $\lambda_1$ and $\lambda_2$ are the solutions to

$$\sum_{i \in A_D_j} \frac{w_{ij0}^*}{1 + \lambda_1(y_i - t_{1j}) + \lambda_2(y_i^2 - t_{2j})} \begin{pmatrix} y_i \\ y_i^2 \end{pmatrix} = \begin{pmatrix} t_{1j} \\ t_{2j} \end{pmatrix}. \quad (4.4.7)$$

In general, a Newton-Raphson method can be used to solve the nonlinear equations (4.4.7). See, for example, Chen, Sitter and Wu (2002).

The existence of the solution is guaranteed by the following two conditions: (i) $(g(x_j, \hat{\beta}), g(x_j, \hat{\beta})^2 + \hat{\sigma}^2 q(x_j, \hat{\beta}))$ is an interior point of the convex hull of $\{(y_i, y_i^2), i \in A_D_j\}$ and (ii)

$$\sum_{i \in A_D_j} w_{ij0}^* \left( \frac{y_i - t_{1j}}{y_i^2 - t_{2j}} \right) (y_i - t_{1j}, y_i^2 - t_{2j})$$

is positive definite. These two conditions should be considered when constructing donor sets. In this paper, two approaches are proposed for constructing donor sets: using observed values and pseudo values. Using observed values can keep the desirable properties of the hot deck method. However, a donor set, $A_{D_j}$, based on observed $y \in A_r$ cannot satisfy two conditions in some situations. For example, when $y_j$ corresponding to the largest or smallest $x_j$ is missing, we cannot make a donor set satisfying the above condition (i). The other approach using pseudo values easily makes donor sets meeting the two conditions even though artificial values are used as imputed values. The choice between the two approaches may depend on practical considerations.
4.5. Some Theory for FDI

When the fractional weights are constructed by satisfying condition (4.3.7), the proposed imputed estimators $\hat{\theta}_{FDI,1n}$ and $\hat{\theta}_{FDI,2n}$ are algebraically equivalent to the DI estimators, $\hat{\theta}_{DI,1n}$ and $\hat{\theta}_{DI,2n}$ respectively. However, the proposed imputed estimator for $\theta_{3N}$ is different from $\hat{\theta}_{DI,3n}$.

It should be possible to show that $\hat{\theta}_{FDI,3n}$ is a consistent estimator of $\theta_{3N}$. The reason that this estimator is consistent is that the weights are constructed so that the expected value of the square of a missing $y$ is matched exactly by the weighted sum of the squared donors. The assumptions required for such a proof would be [C1]-[C5], the other conditions stated in the previous section, and fractional weights satisfying (4.3.7).

It should also be possible to show that the variance of $\hat{\theta}_{FDI,3n}$ is asymptotically equivalent to the variance of $\hat{\theta}(\beta_0)$, where

$$\hat{\theta}(\beta_0) = \sum_{i \in A} w_i \left[ R_i \{ y_i^2 + C^T h(x_i, \beta_0)(y_i - g(x_i, \beta_0)) \} + (1 - R_i)(g(x_i, \beta_0)^2 + \hat{\sigma}^2(\beta_0)) \right]$$

and

$$C = -N^{-1} \left[ \mathbb{E} \left( \frac{\partial \hat{U}(\beta)}{\partial \beta} | \beta = \beta_0 \right) \right]^{-1} \left[ \mathbb{E} \left( \frac{\partial \hat{\theta}_{FDI,3n}}{\partial \beta} | \beta = \beta_0 \right) \right].$$

The assumptions would be the same as previously stated. The reason there should be asymptotic equivalence is that under the assumptions there should be uniform convergence of $\hat{\theta}$ to $\beta_0$, smoothness of the $g$ function, and weights of appropriate nature. Using the asymptotic equivalence, the variance estimator of $\hat{\theta}(\beta_0)$ can be used for estimating the variance of the FDI estimator.

Since $\hat{\theta}_{FDI,t}$, $t = 1, 2, 3$, are consistent and approximately unbiased for $\theta_{1N}$, $\theta_{2N}$ and $\theta_{3N}$, both the FDI estimators for the population correlation coefficient and variance of $Y$ are consistent and approximately unbiased. Note that the FDI estimators for the population correlation coefficient or variance of $Y$ are differentiable functions of $\theta_{t,N}$. Because these estimators are nonlinear functions of many terms, variance estimation
using linearization and substitution involves very messy derivations. In addition, one of main objectives is to provide a single set of replicate weights for general purpose variance estimation. For variance estimation for our estimators, we consider a replicate variance estimation procedure. Details appear in section 4.6.

4.6. Variance estimation for FDI

We now consider variance estimation for the FDI estimators. It is well known that if we treat imputed values as observed values and apply standard variance estimation formulas, then we may seriously underestimate the true variances. Serious underestimation of the true variance of the imputed estimator can lead to erroneous inferences because the additional variability due to the missing values is not being taken into account. For solving this underestimation problem, various variance estimation methodologies have been suggested. They include multiple imputation (Rubin, 1978), adjusted jackknife variance estimation (Rao and Shao, 1992) and fractional imputation with replication variance estimation (Kim and Fuller, 2004 and Fuller and Kim 2005a), respectively. The fractional imputation procedure and variance estimation procedure given by Kim and Fuller (2004) and Fuller and Kim (2005a) is useful for general purposes since in their construction of replicates only the weights on the imputed values are changed. This means that the replicate weights can be easily used for estimators of smooth function of the response variable mean without recomputing imputed values. When the final user is different than the data provider, it is common practice to include a set of replicate weights in the data set. Fuller and Kim (2005a) point out the advantage of providing a single set of replicate weights. “A single set of replicates can be used for variance estimation for imputed variables, variables observed on all respondents, and under assumptions, for function of the two types of variables.”

Since the adjusted imputation fractions under the proposed procedure are functions of the study variable, it is difficult to derive the variance of estimators. Consequently,
the method of Kim and Fuller (2004) and Kim (2007) may not be applied. Instead, our proposed replication variance estimation procedure matches each replicate estimate for the FDI estimator with each replicate estimate for the imputation estimator using predictive values (i.e., the deterministic estimator). The matching is accomplished through proper adjustment of the replication imputation fractions. It is reasonable to do this because of the algebraic equivalence of the two estimators.

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

\[ \hat{V}_{JK}(\hat{\theta}_n) = \sum_{k=1}^{L} c_k \left( \hat{\theta}^{(k)}_n - \hat{\theta}_n \right)^2, \]  

(4.6.1)

where \( \hat{\theta}^{(k)}_n \) is the \( k \text{th} \) estimate of \( \hat{\theta}_n \), based on the observations included in the \( k \text{th} \) replicate, \( L \) is the number of replicates, and \( c_k \) is a factor associated with replicate \( k \) determined by the replication method. In the \( k \text{th} \) replicate, \( w^{(k)}_i = 0 \) if \( i = k \). When the original estimator \( \hat{\theta}_n \) is a linear estimator, the \( k \text{th} \) replicate estimate of \( \hat{\theta}_n \) can be written

\[ \hat{\theta}^{(k)}_n = \sum_{i \in A} w^{(k)}_i y_i, \]  

(4.6.2)

where \( w^{(k)}_i \) denotes the replicate weight for the \( i \text{th} \) unit of the \( k \text{th} \) replication. Assume that the replication variance estimator is consistent for the variance of \( \hat{\theta}_n \).

Let the \( k \text{th} \) replicate for the FDI estimator, \( \hat{\theta}^{f}_{FDI,1n} \), be

\[ \hat{\theta}^{f}_{FDI,1n} = \sum_{i \in A_r} w^{(k)}_i y_i + \sum_{j \in A_m} \sum_{i \in A_{D_j}} w^{(k)}_j p^{(k)}_{ij} y_i \]

\[ \equiv \sum_{i \in A_r} w^{(k)}_i y_i + \sum_{j \in A_m} w^{(k)}_j g(x_j, \hat{\beta}^{(k)}). \]  

(4.6.3)

In the above, the \( p^{(k)}_{ij} \) probabilities are determined so that

\[ \sum_{i \in A_{D_j}} p^{(k)}_{ij} y_i = g(x_j, \hat{\beta}^{(k)}). \]
The coefficient estimate $\hat{\beta}^{(k)}$ is the solution to

$$
\hat{U}^{(k)}(\beta) \equiv \sum_{i \in A_r} w_i^{(k)} \{y_i - g(x_i, \beta)\} h_i(\beta) = 0. \quad (4.6.4)
$$

Let the $k^{th}$ replicate for the FDI estimator, $\hat{\theta}_{FDI,2n}^{(k)}$, be

$$
\hat{\theta}_{FDI,2n}^{(k)} = \sum_{i \in A_r} w_i^{(k)} x_i y_i + \sum_{j \in A_m} \sum_{i \in A_{Dj}} w_j^{(k)} p_{ij}^{(k)} x_j y_i = \sum_{i \in A_r} w_i^{(k)} y_i + \sum_{j \in A_m} w_j^{(k)} x_j g(x_j, \hat{\beta}^{(k)}). \quad (4.6.5)
$$

Let the $k^{th}$ replicate for the FDI estimator, $\hat{\theta}_{FDI,3n}^{(k)}$, be

$$
\hat{\theta}_{FDI,3n}^{(k)} = \sum_{i \in A_r} w_i^{(k)} y_i^2 + \sum_{j \in A_m} \sum_{i \in A_{Dj}} w_j^{(k)} p_{ij}^{(k)} y_i^2 = \sum_{i \in A_r} w_i^{(k)} y_i^2 + \sum_{j \in A_m} w_j^{(k)} \{g(x_i, \hat{\beta}^{(k)})^2 + \hat{\sigma}^2(\hat{\beta}^{(k)})q(x_j, \hat{\beta}^{(k)})\}. \quad (4.6.6)
$$

In the above formula, $\hat{\beta}^{(k)}$ is the same as before and

$$
\hat{\sigma}^2(\hat{\beta}^{(k)}) = \left( \sum_{i \in A_r} w_i^{(k)} \right)^{-1} \sum_{i \in A_r} w_i^{(k)} \frac{(y_i - g(x_i, \hat{\beta}^{(k)}))^2}{q(x_i, \hat{\beta}^{(k)})}.
$$

The conditions on the replicate imputation fractions, $p_{ij}^{(k)}$, can be written as

$$
\sum_{i \in A_{Dj}} p_{ij}^{(k)} (1, y_i, y_i^2) = \left( 1, g(x_i, \hat{\beta}^{(k)}), g(x_i, \hat{\beta}^{(k)})^2 + \hat{\sigma}^2(\hat{\beta}^{(k)})q(x_j, \hat{\beta}^{(k)}) \right). \quad (4.6.7)
$$

For replication variance estimation, initial fractional weights $w_{ij0}^{(k)}$ are assigned where $w_{ij0}^{(k)}$ is small, but positive when $i$ is a deleted unit for replicate $k$. This reflects the effect of deleting the donor while still allowing calibration adjustment. That is, calibration adjustment typically works better with more donors and nonzero initial fractional weights allow more donors. In our simulation, the initial fractional weight for donor $k$ to missing element $j$ is set at $w_{kj0}^{(k)} = 0.1w_{k0}^{(k)}$ and adjusted weights for holding $\sum_{i \in A_{Dj}} w_{ij0}^{(k)} = 1$ are assigned to other values. The final replicate fractional weights $p_{ij}^{(k)}$ is obtained by
using the procedure of (4.4.6) with \( \hat{\beta}^{(k)} \) replacing \( \hat{\beta} \), \( \hat{\sigma}^{2(k)} \) replacing \( \hat{\sigma}^2 \), and \( w_{ij0}^* \) replacing \( w_{ij0}^* \). This procedure assigns the effect of deleting an element to the weights on the donors. That is, the fractional weights on the deleted donors should be reduced.

Note that a single set of replicate weights constructed by the proposed method where \( p_{ij}^{(k)} \) satisfies (4.6.7) can be used for variance estimation for smooth functions of imputed variables and auxiliary variable \( x \). Using the adjusted replicates, the replicate variance estimator can be written as

\[
\hat{V}_{JK,I} = \sum_{k=1}^{L} c_k \left( \hat{\theta}_{FDI,n}^{I} - \hat{\theta}_{FDI,n}^{I} \right)^2.
\] (4.6.8)

The replicated imputation fractions satisfying (4.6.7) can be computed by the empirical likelihood method. Provided that the variance estimator of the deterministic imputation estimator is consistent, the proposed variance estimator of the FDI estimator is also consistent. The suggested procedure is closely related to an adjusted jackknife method (Rao and Shao, 1992) for deterministically imputed estimators. However, in the construction of replicates, only the weights on the imputed values are changed so that the replicate weights can be easily used for any smooth function of \( y \) without recomputing imputed values.

In closing, we reiterate the steps to construct a single set of replicates.

**STEP 1**. Compute \( \hat{\beta}^{(k)} \) and \( \hat{\sigma}^{2(k)} \).

**STEP 2**. Choose the initial replicate fractional weights \( w_{ij0}^* \).

**STEP 3**. Compute the replicate imputation fractions using \( \hat{\beta}^{(k)} \), \( \hat{\sigma}^{2(k)} \), and \( w_{ij0}^* \) in **STEP 1** and **STEP 2** to satisfy condition (4.6.7).

### 4.7. Monte Carlo Study

This paper conducted two simulation experiments for comparing the performance of these imputation methods.
4.7.1 Imputation and Estimation Procedures

For the imputation mechanism, we used following imputation methods:

1. FNFI Fractional Nearest Neighbor Imputation.

2. FRHDI Fractional Regression Hot Deck Imputation.

3. MI Multiple Imputation.

4. DI Deterministic Imputation.

5. FDI Fractional Deterministic Imputation.

Fractional imputation is a procedure in which multiple donors, say \( M \), are chosen for each recipient. In the simulation experiments, \( M = 5 \) and \( M = 10 \) donors were used per recipient. The value for each donor is given a weight equal to a fraction of the original weight, where the fraction is typically \( 1/M \). For fractional imputation methods, we used nearest neighbor imputation where the distance is defined on the value of \( X \)'s. That is, the \( M \) respondents with the closest values of \( X_j, j \in A_m \), are selected as donors. The fractional weights are set to \( w_{ij}^* = 1/M \). These weights are used as initial fractional weights for FNNI, FRHDI and FDI.

In the case of FRHDI and FDI, the imputation fractions are constructed by the empirical likelihood technique to satisfy their constraints respectively. Under the empirical likelihood technique, a solution is not guaranteed for some unlucky donor sets. In the FRHDI method, an unlucky donor set occurs when the element \( y_j \) is missing for the smallest or largest \( x_j \). In the case of FDI, an unlucky donor set occurs when the convex hull of \( y \) and \( y^2 \) of the donors does not contain the predictive value of \( y_j \) and \( y_j^2 \) for the missing unit. For this unlucky case, the fractional weights are constructed by a regression weighting technique in this simulation.
Another imputation method used is the multiple imputation (MI), where the imputed values are generated using the simple linear regression model with the method described in Schenker and Welsh (1988).

For fractional imputation, the variance estimation methods respectively described in Section 2 through 4 were applied. The variance estimator for MI adopted from Rubin (1987) is

\[ \hat{V}_M = W_M + \left( 1 + \frac{1}{M} \right) B_M, \]  

(4.7.1)

where

\[ (W_M, \hat{\theta}_M) = M^{-1} \sum_{l=1}^{M} (V_{I(l)}, \hat{\theta}_{I(l)}), \]

\[ B_M = (M - 1)^{-1} \sum_{l=1}^{M} (\hat{\theta}_{I(l)} - \hat{\theta}_M)^2, \]

where \( M \) is the number of the multiple imputations, \( V_{I(l)} \) is the complete-data variance estimator applied to the \( t \)th imputation dataset, and \( \hat{\theta}_{I(l)} \) is a version of \( \hat{\theta} \) computed from the \( t \)th imputation dataset. Note that we used the jackknife variance estimator for computing \( V_{I(l)} \) in this simulation. The term \( W_M \) is called the average within imputation data set variance. The term \( B_M \) expresses the variability of estimates across sets of imputations. The term \( 1 + 1/M \) is greater than 1 to account for the fact that a finite number of imputations (\( M \)) is used for each missing value. An infinite number of imputations theoretically would be needed to express the posterior predictive distribution of a missing value.

4.7.2. The First Simulation Experiment: Monte Carlo Study

In order to demonstrate the performance of the proposed estimators, we generate independently \( B = 5,000 \) random samples of size 100 from an infinite population with
four variables: $y_i, x_i, R_i, z_i$. The $y$ variables were generated by the linear regression model

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

(4.7.2)

where $\beta_0 = 3, \beta_1 = 1$, and $x_i$ and $e_i$ are independently generated from the standard normal distribution, respectively. In addition, the variable $z_i$ to define a domain was generated from the uniform $(0, 1)$ distribution and the response indicator variable $R_i$ from a Bernoulli distribution with the response rate $p = 0.65$. The $y_i$ is observed if and only if $R_i = 1$. The $x_i$ and $z_i$ are observed throughout the finite population. The estimator $\hat{\beta}$ is obtained from the estimating equations

$$\hat{U}(\beta) = \sum_{i \in A} w_i R_i \{y_i - (\beta_0 + \beta_1 x_i)\} \left(\frac{1}{x_i}\right) = 0.$$

Under this set up, $\hat{\beta}$ is the MLE under missing data.

In the infinite population, we can consider five parameters that might be of interest. One is the mean of the variable $y$, which is $\beta_0$, or 3. The second is the mean of $y$ in the domain determined by variable $z$. Since $z$ and $y$ are independent, the mean of $y$ in the domain is also $\beta_0$, or 3. The third is the probability that $y$ is less than 2. Any number could be used, but 2 is used in simulations here. Since $y$ has a mean of 3 and a variance of 2 and a normal distribution, the probability is 0.23975. A fourth parameter is the correlation between $x$ and $y$. This value is 0.7071. A fifth parameter is the variance of $y$. The variance of $y$ is 2. By the way, the variance of $y$ in the domain is also 2, because of the independence of $z$ and $y$.

If one were to imagine a finite population being generated from the super population model, then one could be interested in the finite population parameters. Suppose that the finite population of interest has $N = 10,000$ elements. A sample of size $n = 100$ is a sampling rate of 1 percent. In such a situation, it is not uncommon to ignore finite population corrections. As such estimates of the finite population parameters are used
as estimates of the super population parameters as well. If the sampling scheme is simple random sampling from the finite population, then the sample can be considered a random sample (i.i.d.) from the super population as well. Variability of the estimators can be determined from variability in the super population. If one were going to compare the estimates to the finite population values, then one would need to generate a finite population of size $N = 10,000$ and then draw a sample of $n = 100$ from that finite population. If one simply is going to compare to the super population values and use simple random sampling, then one can simply generate i.i.d. samples of size $n = 100$ directly.

The formulas for five finite population parameters that correspond to the super population parameters are listed below:

$$
\bar{Y}_N = \frac{1}{N} \sum_{i=1}^{N} y_i,
$$

$$
\bar{Y}_{d,N} = \frac{\sum_{i=1}^{N} d_i y_i}{\sum_{i=1}^{N} d_i},
$$

$$
\bar{P}_N = \frac{1}{N} \sum_{i=1}^{N} p_i,
$$

$$
\rho_{yx,N} = \left( \frac{\sum_{i=1}^{N} (y_i - \bar{Y}_N)^2 \sum_{i=1}^{N} (x_i - \bar{X}_N)^2}{\sum_{i=1}^{N} (y_i - \bar{Y}_N) (x_i - \bar{X}_N)} \right)^{-1} \sum_{i=1}^{N} (y_i - \bar{Y}_N) (x_i - \bar{X}_N),
$$

and

$$
S_{yy,N} = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y}_N)^2,
$$

where $I(z_i < 0.25) = d_i$ and $I(y_i < 2) = p_i$.

One of the reasons to use imputation in place of weighting is to improve estimates for domains. In practical situation, completed data sets by imputation methods are often used to construct estimates for domains where this domain information was not used for the imputation model. An estimated domain mean for the full sample can be
represented as

\[
\hat{Y}_{d,n} = \left( \sum_{i \in A} w_i d_i \right)^{-1} \sum_{i \in A} w_i d_i y_i.
\]

The fractional imputation estimator of the domain mean \( \hat{Y}_{d,n} \) was calculated as

\[
\hat{Y}_{d,n}^f = \left( \sum_{i \in A} w_i d_i \right)^{-1} \left( \sum_{i \in A} w_i d_i y_i + \sum_{j \in A_m} \sum_{i \in A} w_j p_{ij} d_j y_i \right).
\]

The full sample estimator of proportion \( P_N \) can be written as

\[
\hat{P}_n = \left( \sum_{i \in A} w_i \right)^{-1} \sum_{i \in A} w_i p_i.
\]

The fractional imputation estimator of the proportion \( P_N \) was calculated as

\[
\hat{P}_n^f = \left( \sum_{i \in A} w_i \right)^{-1} \left( \sum_{i \in A} w_i p_i + \sum_{j \in A_m} \sum_{i \in A} w_j p_{ij} p_i \right).
\]

Complete estimators of correlation and variance can be represented as

\[
\hat{\rho}_{yx,n} = \frac{\sum_{i \in A} w_i x_i y_i - \sum_{i \in A} w_i x_i \sum_{i \in A} w_i y_i / N}{\sqrt{\left( \sum_{i \in A} w_i y_i^2 - \left( \sum_{i \in A} w_i y_i \right)^2 / N \right)} \left( \sum_{i \in A} w_i x_i^2 - \left( \sum_{i \in A} w_i x_i \right)^2 / N \right)}
\]

and

\[
\hat{S}_{yy,n} = \frac{1}{N - 1} \left( \sum_{i \in A} w_i y_i^2 - \left( \sum_{i \in A} w_i y_i \right)^2 / N \right).
\]

Under fractional imputation, estimators of the correlation and variance were calculated as

\[
\hat{\rho}_{yx,n}^f = \frac{\hat{X} Y_n^f - \hat{Y}_n^f \hat{X}_n / N}{\sqrt{\left( \hat{Y} Y_n^f - (\hat{Y}_n)^2 / N \right) \left( \hat{X} X_n - (\hat{X}_n)^2 / N \right)}}
\]

and

\[
\hat{S}_{yy,n}^f = \frac{1}{N - 1} \left( \hat{Y} Y_n^f - (\hat{Y}_n)^2 / N \right),
\]
where

\[
\hat{Y}^I_n = \sum_{i \in A_r} w_i x_i y_i + \sum_{j \in A_m} \sum_{i \in A_r} w_j p_{ij} x_j y_i, \\
\hat{Y}_{\hat{Y}}^I_n = \sum_{i \in A_r} w_i y_i^2 + \sum_{j \in A_m} \sum_{i \in A_r} w_j p_{ij} y_i^2, \\
\hat{Y}_n = \sum_{i \in A_r} w_i y_i + \sum_{j \in A_m} \sum_{i \in A_r} w_j p_{ij} y_i, \\
\hat{X}_n = \sum_{i \in A} w_i x_i. \\
\]

Estimates for FRHDI and FDI were computed with using adjusted imputation fractions according to their conditions. Note that FDI estimates of marginal means, domain means, variances, and correlations are not affected by the number of donors when \( M > 1 \). This is true because the imputation fractions are constructed to satisfy the condition (4.3.7). Estimates of \( P_N \), however, are affected by the number of donors. The imputation fractions are not created with constraints on non-smooth function of \( y \) under consideration.

**Point Estimator Results**

The Monte Carlo results for 5,000 samples are given in Table A.1 through A.4. Table A.1 and A.2 show the mean and variance of the point estimates for five parameters using various estimators. The mean is the average of the 5,000 replicates. The variance is the sample variance based on the 5,000 replicates. The sizes of simulation error for mean values in Table A.1 and A.2 are about 0.002, 0.004, 0.0007, 0.001, and 0.004 for \( \hat{Y}_N, \hat{Y}_{d,N}, \rho_{xy,N}, \) and \( S_{yy,N} \), respectively. Calculation of simulation error is discussed in the Appendix to this paper. Generally, simulation error is small relative to the major differences across methods.

Estimates of simulation error based on the 5,000 point estimates of the mean closely
match the theoretical simulation error based on normal theory. For the mean of \( y \), the variance is 2 and the variance of the sample average (based on complete response) is 0.02 when \( n = 100 \). The mean in Table A.1 is the average of 5000 values so its variance is 0.02 divided by 5000. The square root of that value is 0.002. The domain mean on average is based on 25 observations, so its variance is 0.08. The simulation error is the square root of 0.08 divided by 5000, which is 0.004. The estimated proportion has variance according to a binomial experiment with \( n = 100 \) sample size \((.24(.76)/100)\). The square root of this value divided by 5000 is 0.0006. The variance of a sample correlation coefficient is approximately \((1 - r^2)/(n - 2)\). Using \( r = 0.7071 \), the square root of this value divided by 5000 is 0.001. For the variance, its variance is approximately 0.0808. The square root of this divided by 5000 is 0.004. The estimated simulation errors are a little larger in the cases with missing data, which reflects the impact of missing data.

Complete indicates that the complete sample is used for estimation and imputation is not necessary. FNNI is fractional nearest neighbor imputation where nearest neighbors are found in the \( x \)-dimension. MI is multiple imputation using the linear regression model. FRHDI is fractional regression hot deck imputation. DI is deterministic imputation, which in this case is regression imputation. FDI is the proposed method, fractional deterministic imputation.

For the mean of \( y \), all methods are unbiased, which is expected because the missing mechanism was ignorable and the parameter is a linear function of the data. FDI in this case is algebraically equivalent to DI and shares the efficiency of DI. FDI and DI point estimates are a little less variable than the other imputation methods. FDI does not depend on \( M \) due to the algebraic equivalence to DI. For the three methods (FNNI, FRHDI, MI) that depend on \( M \), a large number of imputations leads to a more stable point estimator as can be seen by the lower variance with \( M = 10 \).

For the domain mean parameter, all methods produce on average an unbiased estimate of the parameter. Little difference is seen in the variance of the different estimators.
Estimators that depend on $M$ again do better with $M = 10$ than with $M = 5$. For the domain mean, the DI and FDI estimators are algebraically equivalent. In this simulation, the study variable $y$ is independent of $z$. This independence is considered in the imputation method. As a result, some estimators are doing better than the complete sample estimator in terms of variance.

For the parameter $P_N = Pr(y < 2)$, based on simulations, some estimators can be categorized as less acceptable than others. FNNI tends to underestimate $P_N$. This occurs because the nearest neighbor tends to be closer to the center of the $x$-range, which corresponds to larger $y$-values, than $y = 2$. DI also tends to underestimate $P_N$. Since DI is regression imputation, it could be affected negatively for this parameter due to regression to the mean. FRHDI seems to be unbiased, but more variable than MI or FDI. MI in this simulation has quite stable for this parameter. Thus, FDI or MI with $M = 10$ seem superior than the other methods for this parameter.

For the correlation, the FDI and MI methods perform better than the alternatives under consideration. DI distorts the distribution of the variable by placing all imputed points on the regression line. As a result, correlation is overestimated. In contrast, FNNI underestimates the correlation due to a regression to the mean effect. As with the previous estimators, FRHDI is unbiased but variance is larger than for MI and FDI. MI does better with $M = 10$ than with $M = 5$, but FDI does better yet.

In terms of estimating the variance of $y$, the story based on results in Table A.2 is essentially the same as for the correlation. FNNI and DI are biased downwards and are not recommended. FRHDI and MI do better with $M = 10$ than with $M = 5$. FRHDI has more variability than either MI or FDI. FDI does the best among the unbiased estimators with imputation in terms of variance.

**Variance Estimator Results**

In Tables A.3 and A.4, we compare estimates of variance for three estimators: FRHDI, MI, and FDI. These three methods produced unbiased estimators of the parameter in
Tables A.1 and A.2. Table A.3 and A.4 show the relative biases, t-statistics and relative variances for the variance estimators. The relative bias of the variance estimate is estimated by

$$RB(\%) = \frac{E_B(\hat{V}) - V_B(\hat{\theta}_I)}{V_B(\hat{\theta}_I)} \times 100,$$

where $E_B$ and $V_B$ are the Monte Carlo expectation and variance. The t-statistic for testing $H_0 : E(\hat{V}) = Var(\hat{\theta}_I)$ is

$$t = \sqrt{B} \frac{E_B(\hat{V}) - V_B(\hat{\theta}_I)}{V_B(\hat{V} - (\hat{\theta}_I - E_B(\hat{\theta}_I))^2)}.$$

The relative variance is written as

$$RV(\%) = \frac{V_B(\hat{V})}{V_B(\hat{\theta}_I)^2} \times 100.$$

The relative variance will be large if either the Monte Carlo variance of the variance estimators is large (numerator) or the Monte Carlo variance of the point estimator is small (denominator). The same five parameters are considered. The sizes of simulation error for the relative bias values with $M = 10$ in Table A.3 and Table A.4 are about 0.3%, 0.7%, 0.4%, 0.8%, and 0.7% for $\bar{Y}_N$, $\bar{Y}_{d,N}$, $P_N$, $\rho_{yx,N}$, and $S_{yy,N}$, respectively. The sizes of simulation error for the relative variance values with $M = 10$ are 0.1%, 0.5%, 0.1%, 0.5%, and 0.5%. For $M = 5$, MI variance estimators have the larger simulation errors for the relative bias and relative variance because of a small degrees of freedom. FRHDI variance estimators also have the larger simulation errors since the variance of FRHDI variance estimators is proportional to the weights so that negative weights constructed with an unlucky sample lead to a large variance in this simulation. In the case of FDI, those values are much smaller than those of MI and FRHDI variance estimators. Even though the weight set constructed by the FDI scheme with $M = 5$ have negative weights, the variance of FDI variance estimator is small since the replicate weight set is constructed by matching each replicate estimate to that of the DI estimators. The negative weights
do not affect the variance except the variance of the proportion estimator. Note that the proper size of $M$ should be considered for the existence of the solution in (4.3.7).

For the overall mean, the three methods produce about the same relative bias for the estimate of variance. The t-statistics for the variance of this parameter are all about the same and are not extreme. The relative variances also are not extremely different or large. These results indicate that the three methods do about equally well for estimating the variance of the mean of $y$.

Results for the variance estimator of the variance of the domain mean estimator, however, are different. In particular, MI produces a seriously biased estimator of the variance of the estimator of $\bar{Y}_{d,N}$. This property was first identified by Fay (1992). Additionally, Kim and Fuller (2004) discussed that the $B_M$ of (4.7.1) is a biased estimator of the effect of imputing for missing values so that the MI variance estimator does not reflect the fact that some of the imputed values used in the domain come from observations outside the domain. This bias of MI has little relationship to $M$ and cannot be improved by increasing $M$. Because one of the biggest reasons for imputation over weighting is to improve estimators for small domains, this fact is a serious weakness point of using the MI method. The bias of the multiple variance estimators in survey sampling is discussed by Kim, Brick, Fuller and Kalton (2006) and a bias-adjusted variance estimator is suggested. The relative bias of FRHDI is a little larger than for FDI in this case, and both are much smaller than for MI. FRHDI with $M = 10$ does better in terms of relative bias, t-statistic, and relative variance than with $M = 5$. FDI does slightly better than FRHDI with $M = 10$.

Results for the variance estimator of the variance of the estimator of $P_N$ again show an advantage to FRHDI and FDI over MI. In this simulation, the relative bias of the MI variance estimator of the variance of the estimator of $P_N$ is 18%. For FRHDI and FDI, relative bias is less than 2% and the bias is not statistically significant as determined by the t-statistic. All methods have comparable relative variance, but the issue of
bias suggests a preference for FRHDI or FDI. MI is known to have a small sample bias, which was discussed in Kim (2002). The discussion of FDI above concerned the case when $M = 10$. In the case where $M = 5$, some problems with negative weights were encountered since donor sets constructed by the nearest neighbor criterion are not guaranteed to hold two conditions for the existence of the solution in (4.6.7). The proposed variance can be seriously affected by negative weights. When $M = 10$, the problem seems to have been mitigated since there is a chance to increase a possibility holding two conditions as $M$ increases. In practice, the method of constructing a donor set satisfying the two conditions should be considered.

Table A.4 presents variance estimation results for correlation and variance. When $M = 10$, there are not large differences in the results by method. When $M = 5$, FRHDI show some instability in variance estimates. For both of these parameters, MI and FDI perform about the same, whereas FRHDI in simulation had a slight negative bias.

4.7.2. The Second Simulation Experiment: Monte Carlo Study

In the second simulation, $B = 5,000$ random samples of size 200 were generated from an infinite population with four variables $y_i, x_i, R_i, z_i$. The response $y$ variables were generated by the nonlinear regression model

$$y_i = \frac{\beta_0}{1 + \beta_1 \exp(\beta_2 x_i)} + e_i$$

where $(\beta_0 = 10, \beta_1 = 15, \beta_2 = -3)$, $x_i$ is independently generated from $U(0.1, 4.1)$, and $e_i$ is independently generated from $N(0, 1)$. Note that this model is called the logistic regression model generally used in population studies to relate, for instance, number of species $y$ to time $x$. In addition, the response indicator variable $R_i$ is generated from a Bernoulli distribution with parameter $\eta_i = Pr(R_i | x_i)$,

$$\eta_i = \frac{\exp(\phi_0 + \phi_1 x_i)}{1 + \exp(\phi_0 + \phi_1 x_i)}$$
where $\phi_0 = 3$ and $\phi_1 = -1$. The average response rate in this simulation is 66%. In this model (4.7.3), the corresponding $g(x_i, \beta) = \frac{\beta_0}{1 + \beta_1 \exp(\beta_2 x_i)}$, $q(x_i, \beta) = 1$, and

$$h(x_i, \beta) = \begin{pmatrix} [1 + \beta_1 \exp(\beta_2 x_i)]^{-1} \\ -\beta_0 \exp(\beta_2 x_i)[1 + \beta_1 \exp(\beta_2 x_i)]^{-2} \\ -\beta_0 \beta_1 x_i \exp(\beta_2 x_i)[1 + \beta_1 \exp(\beta_2 x_i)]^{-2} \end{pmatrix}.$$

The estimator $\hat{\beta}$ is obtained from the estimating equations

$$\hat{U}(\beta) = \sum_{i \in A} w_i R_i \{y_i - g(x_i, \beta)\} h(x_i, \beta) = 0.$$

Because these equations are nonlinear, they require solution by numerical optimization. The Gauss-Newton method was used to solve the equation. One example sample is graphed in Figure 1.

Three parameters are considered as targets of estimation in this simulation. The first parameter is the mean of $y$. The equation for the mean of $y$ is nonlinear in $x$, so we cannot produce an exact analytical solution. A Taylor series approximation is one
option for determining an approximate value. Alternatively, simulation can be used. We simulated 10 million values of $y$ from the infinite population and computed the mean. The resulting value as 7.92. The second parameter is the correlation between $x$ and $y$. In the large simulation the correlation was 0.807. The third super population parameter was the variance of $y$. In the large simulation the value computed as 9.82.

Estimators were constructed by the DI scheme and by the FDI scheme with $M = 5$ donors. In the FDI scheme, donor sets are constructed by using pseudo values. The parameters of interest are the mean of $y$, denoted by $\check{Y}_N$, the correlation of $y$ and $x$, denoted by $S_{yx,N}$, and the variance of $y$, denoted by $S_{yy,N}$. Table A.5 presents Monte Carlo means and variances for the complete data, FRHDI, DI, and FDI estimators. The sizes of simulation error for mean values in Table A.5 are about 0.3%, 0.04%, and 1.4% for $\check{Y}_N$, $\rho_{xy,N}$, and $S_{yy,N}$, respectively. The properties of the point estimates are similar to those for simulation 1. The results of variance estimators are shown in Table A.6. The sizes of simulation error for the relative bias values in Table A.6 are about 0.13%, 0.4%, and 0.2% for $\check{Y}_N$, $\rho_{yx,N}$, and $S_{yy,N}$, respectively. The sizes of simulation error for the relative variance values are 0.02%, 0.13%, and 0.04%. The pattern of results of simulation 2 is very similar to those of simulation 1. Although the jackknife variance estimators of DI estimators are to obtain consistent variance estimators for $\rho_{xy,N}$ and $S_{yy}$, it is meaningless since the point estimators are biased. The result supports that the proposed method is also working well in more general situation. Unlike simulation 1, most final fractional weights are positive.

5. Conclusion

In fractional imputation, multiple donors are used for each missing unit and each donor is given a fraction of the weight of the nonrespondent. In the weighting system, the imputation fractions can be modified for combining auxiliary information into the
imputed estimator. Under a semi-parametric model, the Deterministic Imputation (DI) method can be used. In practice, the DI method is used to impute missing values marginally.

Since the Deterministic Imputation (DI) method produces completed data sets in which the distribution of $y$ is compressed, it is not a useful method for general purpose analysis. Based on this approach, we consider a semi-parametric imputation method under the framework of fractional imputation. To that end, we propose a new imputation method (FDI) using empirical likelihood to provide an easy-to-use data set for general purpose and keep the desirable properties of the DI method.

It is shown that the use of FDI with a small number of imputations per nonrespondent can give a fully model efficient estimator of the mean. Estimates of other parameters, such as the population correlation coefficient and the population variance, based on completed data set are reasonable estimates. A computationally efficient variance estimator is given that permits the construction of general purpose replicates for variance estimation. Constructed replicate weights sets can be directly used for variance estimation for imputed variables, variables observed on all respondents, and other variables that are not part of the assumed population model without further computation. Variance estimates given by using the replicate weight set are consistent and much smaller estimated variances than multiple imputation estimators. Using pseudo support values is proposed for avoiding negative fractional weights.

Even though FDI is model-based approach method, the FDI procedure is more robust than the MI procedure against misspecification of the random mechanism by which the data were generated. MI generates the imputed values from the conditional distribution, thus requiring the assumption of the random mechanism. The required assumption for FDI is expectation and variance of the study variable. Often, there is no theory available on checking the random mechanism, but the mean and variance of the study variable may be specified by prior information in practice. However, FDI is not protected for
the failure of the assumed model. For more protection, our procedure can be applied to doubly protected imputation in Scharfstein, Rotnitsky, and Robins (1999), Van der Laan and Robins (2003) and Kim and Park (2006). Future work could investigate this possibility.

The proposed imputation method is to share the advantage of deterministic imputation, stochastic imputation and hot deck imputation method. Thus, our method is a useful tool for general-purpose data analysis under missing data. The proposed imputation method is appealing for handling nonresponse because it moves the burden of dealing with the missing data off of data analysts and on to data producers. Specially, the analyst can compute point and variance estimates of interest with a single weight set. One of biggest advantage of the method proposed is its breadth of applicability. It is applicable to the general regression imputation model and likelihood-based estimation in circumstances where the imputed estimator with existing consistent jackknife variance estimator is consistent retaining some practical advantages as we mentioned.
Table A.1  Monte Carlo means and variances based on 5,000 replicates for point estimators of the mean, the domain mean, and a proportion using various methods under the first simulation conditions. NOTE: Simulation error in parentheses.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Imputation Method</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Complete</td>
<td>3.00 (0.0020)</td>
<td>0.0195 (0.0004)</td>
</tr>
<tr>
<td>$\bar{Y}_N$</td>
<td>FNNI (M=5)</td>
<td>3.00 (0.0022)</td>
<td>0.0245 (0.0005)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=5)</td>
<td>3.00 (0.0023)</td>
<td>0.0253 (0.0005)</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>3.00 (0.0022)</td>
<td>0.0249 (0.0005)</td>
</tr>
<tr>
<td></td>
<td>FNNI (M=10)</td>
<td>3.00 (0.0022)</td>
<td>0.0243 (0.0005)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=10)</td>
<td>3.00 (0.0022)</td>
<td>0.0244 (0.0005)</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>3.00 (0.0022)</td>
<td>0.0242 (0.0005)</td>
</tr>
<tr>
<td></td>
<td>DI</td>
<td>3.00 (0.0022)</td>
<td>0.0238 (0.0005)</td>
</tr>
<tr>
<td></td>
<td>FDI</td>
<td>3.00 (0.0022)</td>
<td>0.0238 (0.0005)</td>
</tr>
<tr>
<td>Domain Mean</td>
<td>Complete</td>
<td>3.00 (0.0040)</td>
<td>0.0795 (0.0016)</td>
</tr>
<tr>
<td>$\bar{Y}_{d,N}$</td>
<td>FNNI (M=5)</td>
<td>3.00 (0.0039)</td>
<td>0.0765 (0.0015)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=5)</td>
<td>3.00 (0.0040)</td>
<td>0.0797 (0.0016)</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>3.00 (0.0039)</td>
<td>0.0787 (0.0016)</td>
</tr>
<tr>
<td></td>
<td>FNNI (M=10)</td>
<td>3.00 (0.0040)</td>
<td>0.0743 (0.0015)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=10)</td>
<td>3.00 (0.0040)</td>
<td>0.0771 (0.0015)</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>3.00 (0.0039)</td>
<td>0.0771 (0.0015)</td>
</tr>
<tr>
<td></td>
<td>DI</td>
<td>3.00 (0.0039)</td>
<td>0.0755 (0.0015)</td>
</tr>
<tr>
<td></td>
<td>FDI</td>
<td>3.00 (0.0039)</td>
<td>0.0755 (0.0015)</td>
</tr>
<tr>
<td>$Pr(Y &lt; 2)$</td>
<td>Complete</td>
<td>0.240 (0.0006)</td>
<td>0.0019 (0.00004)</td>
</tr>
<tr>
<td>$P_N$</td>
<td>FNNI (M=5)</td>
<td>0.239 (0.0007)</td>
<td>0.0025 (0.00005)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=5)</td>
<td>0.240 (0.0007)</td>
<td>0.0026 (0.00005)</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>0.240 (0.0006)</td>
<td>0.0020 (0.00004)</td>
</tr>
<tr>
<td></td>
<td>FNNI (M=10)</td>
<td>0.238 (0.0007)</td>
<td>0.0025 (0.00005)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=10)</td>
<td>0.240 (0.0007)</td>
<td>0.0025 (0.00005)</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>0.240 (0.0006)</td>
<td>0.0020 (0.00004)</td>
</tr>
<tr>
<td></td>
<td>DI</td>
<td>0.216 (0.0007)</td>
<td>0.0023 (0.00005)</td>
</tr>
<tr>
<td></td>
<td>FDI (M=5)</td>
<td>0.240 (0.0009)</td>
<td>0.0040 (0.00008)</td>
</tr>
<tr>
<td></td>
<td>FDI (M=10)</td>
<td>0.241 (0.0007)</td>
<td>0.0024 (0.00005)</td>
</tr>
</tbody>
</table>
Table A.2  Monte Carlo means and variances based on 5,000 replicates for point estimators of correlation and variance using various methods under the first simulation conditions. NOTE: Simulation error in parentheses.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Imputation Method</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{y_2,N}$</td>
<td>Complete</td>
<td>0.70 (0.0007)</td>
<td>0.0027 (0.00005)</td>
</tr>
<tr>
<td></td>
<td>FNNI (M=5)</td>
<td>0.69 (0.0009)</td>
<td>0.0039 (0.00008)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=5)</td>
<td>0.71 (0.0010)</td>
<td>0.0047 (0.00009)</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>0.70 (0.0009)</td>
<td>0.0038 (0.00008)</td>
</tr>
<tr>
<td></td>
<td>FNNI (M=10)</td>
<td>0.68 (0.0009)</td>
<td>0.0039 (0.00008)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=10)</td>
<td>0.71 (0.0009)</td>
<td>0.0040 (0.00008)</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>0.70 (0.0009)</td>
<td>0.0037 (0.00007)</td>
</tr>
<tr>
<td></td>
<td>DI</td>
<td>0.76 (0.0008)</td>
<td>0.0030 (0.00006)</td>
</tr>
<tr>
<td></td>
<td>FDI</td>
<td>0.70 (0.0008)</td>
<td>0.0036 (0.00007)</td>
</tr>
<tr>
<td>Variance of $Y$</td>
<td>Complete</td>
<td>2.00 (0.0041)</td>
<td>0.0829 (0.0017)</td>
</tr>
<tr>
<td></td>
<td>FNNI (M=5)</td>
<td>1.93 (0.0045)</td>
<td>0.1046 (0.0021)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=5)</td>
<td>1.99 (0.0049)</td>
<td>0.1215 (0.0024)</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>2.00 (0.0047)</td>
<td>0.1126 (0.0023)</td>
</tr>
<tr>
<td></td>
<td>FNNI (M=10)</td>
<td>1.91 (0.0045)</td>
<td>0.0998 (0.0020)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=10)</td>
<td>1.99 (0.0047)</td>
<td>0.1117 (0.0022)</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>2.00 (0.0047)</td>
<td>0.1095 (0.0022)</td>
</tr>
<tr>
<td></td>
<td>DI</td>
<td>1.69 (0.0043)</td>
<td>0.0922 (0.0018)</td>
</tr>
<tr>
<td></td>
<td>FDI</td>
<td>2.00 (0.0045)</td>
<td>0.1053 (0.0021)</td>
</tr>
</tbody>
</table>
Table A.3  Relative biases, t-statistics, and relative variance for variance estimators of the mean, the domain mean, and a proportion using various methods under the first simulation conditions. NOTE: Simulation error in parentheses.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Imputation Method</th>
<th>RB(%)</th>
<th>t-statistic</th>
<th>RV(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>FRHDI (M=5)</td>
<td>3.27 (0.42)</td>
<td>1.62</td>
<td>8.89 (0.18)</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>2.75 (0.35)</td>
<td>1.35</td>
<td>6.11 (0.12)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=10)</td>
<td>2.35 (0.26)</td>
<td>1.17</td>
<td>3.26 (0.07)</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>3.29 (0.29)</td>
<td>1.65</td>
<td>4.23 (0.08)</td>
</tr>
<tr>
<td></td>
<td>FDI</td>
<td>2.94 (0.23)</td>
<td>1.47</td>
<td>2.74 (0.05)</td>
</tr>
<tr>
<td>Domain Mean</td>
<td>FRHDI (M=5)</td>
<td>11.95 (0.82)</td>
<td>5.69</td>
<td>33.47 (0.67)</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>31.02 (0.72)</td>
<td>14.77</td>
<td>25.74 (0.51)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=10)</td>
<td>8.68 (0.61)</td>
<td>4.14</td>
<td>18.48 (0.37)</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>32.11 (0.69)</td>
<td>15.32</td>
<td>24.03 (0.48)</td>
</tr>
<tr>
<td></td>
<td>FDI</td>
<td>7.16 (0.58)</td>
<td>3.45</td>
<td>16.54 (0.33)</td>
</tr>
<tr>
<td>Pr(Y &lt; 2)</td>
<td>FRHDI (M=5)</td>
<td>0.66 (0.42)</td>
<td>0.32</td>
<td>8.87 (0.18)</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>18.31 (0.37)</td>
<td>8.95</td>
<td>6.91 (0.14)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=10)</td>
<td>-0.91 (0.24)</td>
<td>-0.44</td>
<td>2.96 (0.06)</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>17.89 (0.30)</td>
<td>8.81</td>
<td>4.20 (0.08)</td>
</tr>
<tr>
<td></td>
<td>FDI(M=5)</td>
<td>1.36 (1.34)</td>
<td>0.73</td>
<td>89.21 (1.78)</td>
</tr>
<tr>
<td></td>
<td>FDI(M=10)</td>
<td>1.76 (0.30)</td>
<td>0.88</td>
<td>4.80 (0.09)</td>
</tr>
</tbody>
</table>
Table A.4  Relative biases, t-statistics, and relative variance for variance estimators of correlation and variance using various methods under the first simulation conditions. NOTE: Simulation error in parentheses.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Imputation Method</th>
<th>RB(%)</th>
<th>t-statistic</th>
<th>RV(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation $\rho_{yx,N}$</td>
<td>FRHDI (M=5)</td>
<td>5.82 (2.31)</td>
<td>2.63</td>
<td>267.46 (5.35)</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>0.29 (0.72)</td>
<td>0.14</td>
<td>25.74 (0.51)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=10)</td>
<td>1.25 (0.84)</td>
<td>0.59</td>
<td>35.69 (0.71)</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>2.00 (0.64)</td>
<td>0.97</td>
<td>20.60 (0.41)</td>
</tr>
<tr>
<td></td>
<td>FDI</td>
<td>2.48 (0.64)</td>
<td>1.21</td>
<td>20.59 (0.41)</td>
</tr>
<tr>
<td>Variance of Y $S_{yy,N}$</td>
<td>FRHDI (M=5)</td>
<td>-1.21 (1.24)</td>
<td>-0.59</td>
<td>76.99 (1.54)</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>3.14 (0.71)</td>
<td>1.58</td>
<td>24.92 (0.50)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=10)</td>
<td>-0.54 (0.68)</td>
<td>-0.28</td>
<td>22.86 (0.46)</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>3.17 (0.64)</td>
<td>1.62</td>
<td>20.31 (0.41)</td>
</tr>
<tr>
<td></td>
<td>FDI</td>
<td>2.17 (0.59)</td>
<td>1.10</td>
<td>17.64 (0.35)</td>
</tr>
</tbody>
</table>
Table A.5  Monte Carlo means and variances based on 5,000 replicates for point estimators of the mean, correlation, and variance using various methods under the second simulation conditions. NOTE: Simulation error in parentheses.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Imputation Method</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Complete</td>
<td>7.92 (0.0031)</td>
<td>0.0496 (0.0010)</td>
</tr>
<tr>
<td>$\bar{Y}_N$</td>
<td>FNNI (M=5)</td>
<td>7.92 (0.0033)</td>
<td>0.0542 (0.0011)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=5)</td>
<td>7.92 (0.0033)</td>
<td>0.0555 (0.0011)</td>
</tr>
<tr>
<td></td>
<td>FNNI (M=10)</td>
<td>7.92 (0.0033)</td>
<td>0.0542 (0.0011)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=10)</td>
<td>7.92 (0.0033)</td>
<td>0.0546 (0.0011)</td>
</tr>
<tr>
<td></td>
<td>DI</td>
<td>7.92 (0.0033)</td>
<td>0.0532 (0.0011)</td>
</tr>
<tr>
<td></td>
<td>FDI</td>
<td>7.92 (0.0033)</td>
<td>0.0532 (0.0011)</td>
</tr>
<tr>
<td>Correlation</td>
<td>Complete</td>
<td>0.80 (0.0003)</td>
<td>0.0004 (0.000010)</td>
</tr>
<tr>
<td>$\rho_{yx,N}$</td>
<td>FNNI (M=5)</td>
<td>0.81 (0.0003)</td>
<td>0.0006 (0.000012)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=5)</td>
<td>0.80 (0.0004)</td>
<td>0.0008 (0.000016)</td>
</tr>
<tr>
<td></td>
<td>FNNI (M=10)</td>
<td>0.81 (0.0004)</td>
<td>0.0006 (0.000012)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=10)</td>
<td>0.80 (0.0004)</td>
<td>0.0007 (0.000014)</td>
</tr>
<tr>
<td></td>
<td>DI</td>
<td>0.82 (0.0003)</td>
<td>0.0004 (0.000010)</td>
</tr>
<tr>
<td></td>
<td>FDI</td>
<td>0.80 (0.0003)</td>
<td>0.0004 (0.000010)</td>
</tr>
<tr>
<td>Variance of Y</td>
<td>Complete</td>
<td>9.82 (0.0137)</td>
<td>0.9405 (0.0190)</td>
</tr>
<tr>
<td>$S_{yy,N}$</td>
<td>FNNI (M=5)</td>
<td>9.76 (0.0142)</td>
<td>1.0009 (0.0200)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=5)</td>
<td>9.80 (0.0143)</td>
<td>1.0280 (0.0206)</td>
</tr>
<tr>
<td></td>
<td>FNNI (M=10)</td>
<td>9.76 (0.0141)</td>
<td>1.0005 (0.0200)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=10)</td>
<td>9.80 (0.0142)</td>
<td>1.0064 (0.0201)</td>
</tr>
<tr>
<td></td>
<td>DI</td>
<td>9.43 (0.0141)</td>
<td>0.9901 (0.0198)</td>
</tr>
<tr>
<td></td>
<td>FDI</td>
<td>9.80 (0.0140)</td>
<td>0.9848 (0.0197)</td>
</tr>
</tbody>
</table>
Table A.6 Relative biases, t-statistics, and relative variance for variance estimators of the mean, correlation and variance using various methods under the second simulation conditions. NOTE: Simulation error in parentheses.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Imputation Method</th>
<th>RB(%)</th>
<th>t-statistic</th>
<th>RV(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean ( \hat{Y}_N )</td>
<td>FNNI (M=5)</td>
<td>-1.55 (0.01)</td>
<td>-0.76</td>
<td>0.99 (0.02)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=5)</td>
<td>-1.31 (0.03)</td>
<td>-0.64</td>
<td>3.85 (0.08)</td>
</tr>
<tr>
<td></td>
<td>FNNI (M=10)</td>
<td>-1.81 (0.01)</td>
<td>-0.89</td>
<td>0.95 (0.02)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=10)</td>
<td>-1.90 (0.02)</td>
<td>-0.93</td>
<td>1.19 (0.02)</td>
</tr>
<tr>
<td></td>
<td>DI</td>
<td>-1.29 (0.01)</td>
<td>-0.63</td>
<td>0.92 (0.02)</td>
</tr>
<tr>
<td></td>
<td>FDI</td>
<td>-1.29 (0.01)</td>
<td>-0.63</td>
<td>0.92 (0.02)</td>
</tr>
<tr>
<td>Correlation ( \rho_{yx,N} )</td>
<td>FNNI (M=5)</td>
<td>7.30 (0.06)</td>
<td>3.63</td>
<td>19.16 (0.38)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=5)</td>
<td>5.52 (0.23)</td>
<td>1.86</td>
<td>260.64 (5.21)</td>
</tr>
<tr>
<td></td>
<td>FNNI (M=10)</td>
<td>7.44 (0.05)</td>
<td>3.71</td>
<td>13.52 (0.27)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=10)</td>
<td>6.70 (0.08)</td>
<td>3.12</td>
<td>35.40 (0.71)</td>
</tr>
<tr>
<td></td>
<td>DI</td>
<td>5.90 (0.04)</td>
<td>2.95</td>
<td>6.60 (0.13)</td>
</tr>
<tr>
<td></td>
<td>FDI</td>
<td>5.94 (0.04)</td>
<td>2.96</td>
<td>6.99 (0.14)</td>
</tr>
<tr>
<td>Variance of ( Y )</td>
<td>FNNI (M=5)</td>
<td>3.74 (0.02)</td>
<td>1.86</td>
<td>2.60 (0.05)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=5)</td>
<td>4.75 (0.04)</td>
<td>2.36</td>
<td>7.34 (0.15)</td>
</tr>
<tr>
<td></td>
<td>FNNI (M=10)</td>
<td>3.53 (0.02)</td>
<td>1.74</td>
<td>2.49 (0.05)</td>
</tr>
<tr>
<td></td>
<td>FRHDI (M=10)</td>
<td>4.44 (0.02)</td>
<td>2.21</td>
<td>2.94 (0.06)</td>
</tr>
<tr>
<td></td>
<td>DI</td>
<td>3.29 (0.02)</td>
<td>1.63</td>
<td>2.16 (0.04)</td>
</tr>
<tr>
<td></td>
<td>FDI</td>
<td>3.30 (0.02)</td>
<td>1.64</td>
<td>2.23 (0.04)</td>
</tr>
</tbody>
</table>
Appendix

It is possible to compute the variability in results for the various parameters due to the finite number of simulation replications.

A. Simulation error of point estimator

Let \( \hat{\theta}_i, i = 1, 2, \ldots, B \), be independent, identically distributed samples from a distribution \( G_1(\mu, \sigma^2) \). Then by the central limit theorem, \( E_B(\hat{\theta}) = \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i \) converges in distribution to \( N(\mu, \sigma^2/B) \).

By the weak law of large number and the continuous mapping theorem, we have

\[
V_B(\hat{\theta}) = \frac{1}{B-1} \sum_{i=1}^{B} (\hat{\theta}_i - E_B(\hat{\theta}))^2 = \sigma^2 + o_p(1).
\]

Then, \( \sigma^2 \) is consistently estimated by \( V_B(\hat{\theta}) \). Thus, the simulation error for the Monte Carlo mean of point estimator is consistently estimated by \( \sqrt{V_B(\hat{\theta})/B} \).

For the simulation error for the Monte Carlo variance of a point estimator, we assume that

\[
(B - 1) \frac{V_B(\hat{\theta})}{\sigma^2} \sim \chi^2_{(B-1)}.
\]

By properties of the \( \chi^2 \) distribution, the variance of \( V_B(\hat{\theta}) \) is \( 2\sigma^4/(B-1) \). Thus, the simulation error can be estimated by \( \sqrt{2V_B(\hat{\theta})^2/(B - 1)} \).

B. Simulation error of variance estimator

Let \( \hat{V}_i, i = 1, 2, \ldots, B \), be independent, identically distributed samples from a distribution \( G_2(\sigma^2, V^*) \). By the central limit theorem, we have

\[
E_B(\hat{V}) \sim N \left( \sigma^2, \frac{V^*}{B} \right)
\]

and further more

\[
E_B(\hat{V}) - \sigma^2 \sigma^2 \sim N \left( 0, \frac{V^*}{\sigma^4B} \right).
\]
By Slutsky’s theorem and the continuous mapping theorem,

\[ \frac{E_B(\hat{V}) - V_B(\hat{\theta})}{V_B(\hat{\theta})} \sim N\left(0, \frac{V^*}{\sigma^4 B}\right). \]

\(V^*\) and \(\sigma^2\) are consistently estimated by \(V_B(\hat{V})\) and \(V_B(\hat{\theta})\). Thus, the simulation error of the relative bias of the variance estimator is consistently estimated by \(\sqrt{V_B(\hat{V})/(V_B(\hat{\theta})^2 B)}\).

For the simulation error for the relative variance of variance estimator, we may assume that

\[ (B - 1) \frac{V_B(\hat{V})}{V^*} \sim \chi^2_{(B - 1)} \]

By the continuous mapping theorem and Slutsky’s theorem, we have

\[ (B - 1) \frac{V_B(\hat{V})}{V_B(\hat{\theta})^2} \sim \frac{V^*}{\sigma^4} \chi^2_{(B - 1)} \]

By properties of the \(\chi^2\) distribution, the variance of \(V_B(\hat{V})/V_B(\hat{\theta})^2\) is \(2V^*/((B - 1)\sigma^8)\). Thus, the simulation error of the relative variance of the variance estimator is consistently estimated by \(\sqrt{2V_B(\hat{V})^2/((B - 1)V_B(\hat{\theta})^4)}\)

C. t-statistic rationale

Write

\[ E_B(\hat{V}) - V_B(\hat{\theta}) = \frac{1}{B} \sum_{i=1}^{B} \hat{V}_i + \frac{1}{B - 1} \sum_{i=1}^{B} (\hat{\theta}_i - E_B(\hat{\theta}))^2 \]

\[ \approx \frac{1}{B} \sum_{i=1}^{B} (\hat{V}_i - (\hat{\theta}_i - E_B(\hat{\theta}))^2) \]

\[ = E_B(\hat{V}_i - (\hat{\theta}_i - E_B(\hat{\theta}))^2) \]

Then, by the central limit theorem,

\[ \frac{E_B(\hat{V}_i - (\hat{\theta}_i - E_B(\hat{\theta}))^2) - (E_B(\hat{V}) - V(\hat{\theta}))}{\sqrt{V(E_B(\hat{V}_i - (\hat{\theta}_i - E_B(\hat{\theta}))^2))}} \]
converges in distribution to $N(0,1)$. The variance term in the denominator can be consistently estimated by
\[ \frac{1}{B} V_B(\hat{V} - (\hat{\theta} - E_B(\hat{\theta}))^2). \]
Thus, under $H_0 : E(\hat{V}) = V(\hat{\theta})$, the t-statistic can be defined as
\[ t = \frac{\sqrt{B} \frac{E_B(\hat{V}) - V_B(\hat{\theta})}{\sqrt{V_B(\hat{V}) - (\hat{\theta}_i - E_B(\hat{\theta}))^2}}. \]
Additionally, Slutsky’s theorem gives that the sequence of t-statistics converges in distribution to $N(0,1)$ as $b$ goes to infinity.
EXTENSION OF FRACTIONAL IMPUTATION TO
GENERAL MISSINGNESS PATTERNS USING MAXIMUM
LIKELIHOOD

A paper appeared as a proceedings paper
of the 2008 Joint Statistical Meetings.

Minhui Paik and Michael D. Larsen

Abstract

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

Key Words: EM algorithm; Missing data; Multivariate normal; Replicate variance
estimation; Superpopulation model.
1. The EM Algorithm for Missing Data

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

The idea of the EM algorithm is very closely related to the intuitive idea of cycling between filling in missing values and estimating the parameters. The EM algorithm basically is an iterative method in which the missing values are filled in given current estimates of the parameters, and then the parameters are estimated given the filled in missing values. The technical details are a little more involved. First, imagine a set of complete data that is a superset of the observed data. Second, propose a statistical model for the complete data. The complete data model along with assumptions about why the data are missing (the missingness mechanism) and the observed data determines the observed data likelihood. Third, determine the form of the complete data log likelihood. EM alternates between taking the expectation of the complete data log likelihood given current values of parameters and maximizing the complete data log likelihood for parameters given the expected values. It is necessary to take the expected value of all
functions of the missing data present in the complete data log likelihood. When the complete data log likelihood is linear in the data, then the E-step is a simple expectation. In other cases, the expectation is more complicated but the idea is essentially the same.

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

2. Fractional Imputation and Maximum Likelihood

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

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

3. Notation and Model

Consider a finite population $U = \{1, 2, \ldots, N\}$ with $p$ variables potentially recorded for each subject. For the $i^{th}$ element of the finite population, $y_i = (y_{i1}, \ldots, y_{ip})$, are the values of the $p$ variables. We assume that the finite population is a random sample from a superpopulation model. In this case, we assume the superpopulation model is the $p$-variate normal distribution with mean $\mu$ and covariance matrix $\Sigma = (\sigma_{jk})$. For $i = 1, \ldots, n$, independently and identically distributed,

$$y_i \sim N_p(\mu, \Sigma).$$

Let $Y_N = (y_1, y_2, \ldots, y_N)'$ be the finite population of size $N$. The interesting parameters are the finite population mean of each variable and the finite population covariance between two variables. There are defined as follows:

$$\bar{Y}_{j,U} = \frac{1}{N} \sum_{i \in U} y_{ij}$$

and

$$S_{jk,U} = \frac{1}{N-1} \sum_{i \in U} (y_{ij} - \bar{Y}_{j,U})(y_{ik} - \bar{Y}_{k,U}).$$

For simplicity, the shorter notation $Y_j$ and $S_{jk}$ are used in this section. Note that $S_{jk}$ can be expressed as functions of the population means

$$\bar{Y}_j = N^{-1} \sum_{i \in U} y_{ij}, \quad \bar{Y}_k = N^{-1} \sum_{i \in U} y_{ik},$$

$$\bar{Y}_{jj} = N^{-1} \sum_{i \in U} y_{ij}^2, \quad \bar{Y}_{kk} = N^{-1} \sum_{i \in U} y_{ik}^2,$$

and $\bar{Y}_{jk} = N^{-1} \sum_{i \in U} y_{ij} y_{ik}$

$$= N^{-1} \sum_{i \in U} y_{ij}^2, \quad \bar{Y}_{kk} = N^{-1} \sum_{i \in U} y_{ik}^2, \quad \bar{Y}_{jk} = N^{-1} \sum_{i \in U} y_{ij} y_{ik}$$
as
\[ S_{jk} = \frac{N}{N-1} (\bar{Y}_{jk} - \bar{Y}_j \bar{Y}_k). \]  \hfill (3.4)

If data were observed on the complete sample, estimators of \( \bar{Y}_j \) and \( \bar{Y}_{jk} \) based on the sample \( A \) with size \( n \) are
\[ \bar{y}_{j,n} = \sum_{i \in A} w_i y_{ij} \] and
\[ \bar{y}_{jk,n} = \sum_{i \in A} w_i y_{ij} y_{ik}, \] \hfill (3.5)
where \( w_i = N^{-1} \pi_i^{-1} \), \( \pi_i = P(i \in A) \) is the probability that unit \( i \) is in the sample, and \( A \) is the set of indices in the sample.

An estimator of \( S_{jk} \) is constructed as follows:
\[ s_{jk,n} = \frac{N}{N-1} (\bar{y}_{jk,n} - \bar{y}_{j,n} \bar{y}_{k,n}), \] \hfill (3.6)
where \( \bar{y}_{jk,n}, \bar{y}_{j,n}, \) and \( \bar{y}_{k,n} \) are sample estimates of the corresponding population means. For simplicity, the shorter notation \( y_j, y_{jk} \) and \( s_{jk} \) are used in the remainder of this section. By the definition of \( w_i \), we have
\[ E(\bar{y}_j | F_N) = \bar{Y}_j, \]
\[ E(\bar{y}_{jj} | F_N) = \bar{Y}_{jj}, \] \hfill (3.7)
\[ E(\bar{y}_{jk} | F_N) = \bar{Y}_{jk}, \] and
\[ E(s_{jk} | F_N) = S_{jk}, \]
where \( F_N = \{y_1, \ldots, y_N\} \).

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

**Lemma 1.** Under the superpopulation model and an ignorable sampling mechanism, \( \bar{y}_j \) and \( s_{jk} \) are consistent estimators for the corresponding finite population parameters.
4. Missing Data and the Proposed Method

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

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

To achieve our objective, the imputed values $y_{ij}^*$ for subject $i$ on variable $j$ have to
satisfy the following conditions. We define the response indicator variable of \( y_{ij} \) by

\[
R_{ij} = \begin{cases} 
1 & j^{th} \text{ variable is observed for unit } i \\
0 & j^{th} \text{ variable is not observed for unit } i.
\end{cases}
\]

For \( R_{ij} = 0, R_{ik} = 0, \) and \( R_{is} = 1 \) for \( s \neq j, k, \)

\[
y^*_i = E(y_{ij}|y_{obs,i}, \hat{\theta}), \\
y^*_2 = E(y^2_{ij}|y_{obs,i}, \hat{\theta}), \text{ and } \\
y^*_j y^*_k = E(y_{ij} y_{ik}|y_{obs,i}, \hat{\theta}),
\]

(4.1)

where \( y_{obs,i} \) denotes the set of variables observed for unit \( i \) and \( \hat{\theta} = (\hat{\mu}, \hat{\Sigma}) \) are ML estimates, which possibly are obtained by iterative methods.

Since a single imputed donor value can not satisfy the above conditions, our approach is to use multiple donors and assign adjusted fractional weights to the donors in order to satisfy the conditions. The imputed values based on several donors and fractional weights can be defined as follows:

\[
y^*_i = \sum_{t \in A_{D,i}} w^*_i y_{jt},
\]

(4.2)

where \( A_{D,i} \) is the donor set of indices for missing unit \( i \). Note that this donor set for unit \( i \) can be constructed by a systematic sampling method from available donors sorted in some manner, simple random sampling without replacement from the observed cases, or selection of donors using some nearest neighbor method. If it is not important to use observed values as imputed values, the imputed values can be generated from the conditional distribution given by observed cases \( y_{obs} \) and ML estimates. Then, the proposed method consists of finding the fractional weights satisfying the following constraints. Two cases can be considered. First, for \( R_{ij} = 0 \) and \( R_{ik} = 1, j \neq k, \)

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

(4.3)
Second, for $R_{ij} = 0$, $R_{ik} = 0$ and $R_{is} = 1$ for $s \neq j, k$, the constraints are

$$
\sum_{t \in A_{D,i}} w_t^* \left(1, y_{tj}, y_{tk}, y_{tj}^2, y_{tk}^2\right) = \left(1, E(y_{tj}|y_{obs,i}, \hat{\theta}), E(y_{tk}|y_{obs,i}, \hat{\theta}), E(y_{tj}^2|y_{obs,i}, \hat{\theta}), E(y_{tk}^2|y_{obs,i}, \hat{\theta})\right).
$$

(4.4)

We can use a regression weighting technique or an empirical likelihood technique to find a solution to (4.3) and (4.5). To avoid the chance of extreme weights, including possibly negative weights, the nonnegative fractional weights method of Paik and Larsen (2007) or a modified Newton-Raphson method as in Chen, Sitter, and Wu (2002) can be used to solve the constraints. The size of donor sets for missing values do not necessarily need to be very large in order to do this in general.

5. An Example: Trivariate Normal Sample with Missing Data on Two Variables

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

The constraints (4.3) and (4.5) in this situation can be expressed as follows:
Similarly, we can calculate other conditional expectations:

For $i \in A_{mr}$,
\[
\sum_{t \in A_{D,i}} w^*_t (1, y_t, y^2_t) = \left(1, E(y_t|z_i, x_i, \hat{\theta}), E(y^2_t|z_i, x_i, \hat{\theta})\right).
\]

For $i \in A_{rm}$,
\[
\sum_{t \in A_{D,i}} w^*_t (1, z_t, z^2_t) = \left(1, E(z_t|y_i, x_i, \hat{\theta}), E(z^2_t|y_i, x_i, \hat{\theta})\right).
\]

For $i \in A_{mm}$,
\[
\sum_{t \in A_{D,i}} w^*_t (1, y_t, z_t, y_t z_t, y^2_t, z^2_t)
\]
\[
= \left(1, E(y_t|x_i, \hat{\theta}), E(z_t|x_i, \hat{\theta}), E(y_t z_t|x_i, \hat{\theta}), E(y^2_t|x_i, \hat{\theta}), E(z^2_t|x_i, \hat{\theta})\right).
\]

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

\[
E(y_t|z_i, x_i, \hat{\theta}) = \hat{\mu}_y + (\hat{\sigma}_{yz}, \hat{\sigma}_{yx}) \begin{pmatrix} \hat{\sigma}_{xx} & \hat{\sigma}_{zx} \\ \hat{\sigma}_{zx} & \hat{\sigma}_{xx} \end{pmatrix}^{-1} \begin{pmatrix} z_i - \hat{\mu}_z \\ x_i - \hat{\mu}_x \end{pmatrix},
\]

\[
V(y_t|z_i, x_i, \hat{\theta}) = \hat{\sigma}_{yy} - (\hat{\sigma}_{yz}, \hat{\sigma}_{yx}) \begin{pmatrix} \hat{\sigma}_{xx} & \hat{\sigma}_{zx} \\ \hat{\sigma}_{zx} & \hat{\sigma}_{xx} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\sigma}_{yz} \\ \hat{\sigma}_{yx} \end{pmatrix},
\]

\[
E(y_t|x_i, \hat{\theta}) = \hat{\mu}_y + \hat{\sigma}_{yx} \hat{\sigma}_{xx}^{-1} (x_i - \hat{\mu}_x),
\]

\[
V(y_t|x_i, \hat{\theta}) = \hat{\sigma}_{yy} - \hat{\sigma}_{yx}^2 \hat{\sigma}_{xx}^{-1}, \text{ and}
\]

\[
COV(y_t, z_t|x_i, \hat{\theta}) = \hat{\sigma}_{yz} - \hat{\sigma}_{yx} \hat{\sigma}_{xx}^{-1} \hat{\sigma}_{xx}.
\]

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

The proposed imputed estimators of population means in (3.3) are defined respectively as

\[
\bar{y}_{l,j} = \sum_{i \in A} \sum_{t \in A_{D,i}} w_i w^*_t y_{lj} \quad \text{and} \quad (5.2)
\]

\[
\bar{y}_{l,j,k} = \sum_{i \in A} \sum_{t \in A_{D,i}} w_i w^*_t y_{lj} y_{tk},
\]
where the following values are occur when $R_{ij} = 1$: $w_{it} = 0$ when $i \neq t$, $A_{D,i}$ has only one unit $(y_{ij})$, and $w_{ii} = 1$. In addition, the imputed estimators of population covariance in (3.4) can be written as

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

(5.3)

6. A Theoretical Result

**Theorem 2.** Suppose that the maximum likelihood estimator $\hat{\theta}$ of $\theta$ is available and, under some regularity conditions, $\hat{\theta}$ is a consistent estimator of $\theta$. Then the imputed estimators in (5.2) and (5.3) based on the fractional imputation estimators described previously in this chapter are approximately unbiased estimators for finite population parameters under the superpopulation model in (3.1) and an ignorable missing data mechanism.

**Proof.** We consider only the trivariate normal case in the example. Results for other situations should also be possible. Define $R_{iy} = 1$ if $y_i$ is observed and $R_{iy} = 0$ if $y_i$ is a nonrespondent and $R_{iz} = 1$ if $z_i$ is observed and $R_{iz} = 0$ if $z_i$ is a nonrespondent. We assume that the definition of the response indicator can be extended to the entire population. That is, the indicators tell us whether an individual would respond if sampled
from the population. First,

\[
E(\bar{y}_I - \bar{Y}_N) = E \left( E(\bar{y}_I - \bar{Y}_N|F_N) \right)
\]

\[
= E \left[ E \left( \sum_{i \in A} \sum_{t \in A_{D,i}} w_i (1 - R_{iy}) w_t^* y_t + \sum_{i \in A} w_i R_{iy} y_i - \bar{Y}_N|F_N \right) \right]
\]

\[
= E \left[ E \left( \sum_{i \in A} w_i (1 - R_{iy}) R_{iz} E(y_i|z_i, x_i, \hat{\theta}) + \sum_{i \in A} w_i (1 - R_{iy})(1 - R_{iz}) E(y_i|x_i, \hat{\theta}) \right.ight.
\]

\[
+ \sum_{i \in A} w_i R_{iy} y_i - \bar{Y}_N|F_N \left. \right) \right]
\]

\[
= E \left[ E \sum_{i \in U} (1 - R_{iy}) R_{iz} E(y_i|z_i, x_i, \hat{\theta}) + \sum_{i \in U} (1 - R_{iy})(1 - R_{iz}) E(y_i|x_i, \hat{\theta}) \right.
\]

\[
- \sum_{i \in U} (1 - R_{iy}) y_i \left. \right]\]

\[
\approx E \left[ \sum_{i \in U} (1 - R_{iy}) R_{iz} E(y_i|z_i, x_i, \hat{\theta}) + \sum_{i \in U} (1 - R_{iy})(1 - R_{iz}) E(y_i|x_i, \theta) \right. \]

\[
- \sum_{i \in U} (1 - R_{iy}) y_i \left. \right]\]

\[
= E \left[ \sum_{i \in U} (1 - R_{iy}) R_{iz} y_i + \sum_{i \in U} (1 - R_{iy})(1 - R_{iz}) y_i - \sum_{i \in U} (1 - R_{iy}) y_i \right] = 0
\]

In the third line, the terms in \( E(\bar{y}_I|F_N) \) and \( \bar{Y}_N \) with observed values of \( y \) cancel. The approximation equation in the fourth line is established by the fact that \( E(y_i|z_i, x_i, \hat{\theta}) \) and \( E(y_i|x_i, \hat{\theta}) \) are differentiable functions of the MLE estimator. The reference distribution of \( E(\cdot) \) is the joint distribution of the superpopulation model, the sampling distribution and the response model. Under the assumption of MAR, the last equation is true. Similarly, \( E(\bar{z}_I - \bar{Z}_N) \approx 0 \).
Next,

\[ E(\bar{y}y_I - \bar{YY}_N) = E \left( E(\bar{y}y_I - \bar{YY}_N|F_N) \right) \]

\[ = E \left[ E \left( \sum_{i \in A} \sum_{t \in A_{D,i}} w_i (1 - R_{iy}) w_t^* y_t^2 + \sum_{i \in A} w_i R_{iy} y_i^2 - \bar{YY}_N|F_N \right) \right] \]

\[ = E \left[ E \left( \sum_{i \in A} w_i (1 - R_{iy}) R_{iz} E(y_i^2|z_i, x_i, \hat{\theta}) + \sum_{i \in A} w_i (1 - R_{iy})(1 - R_{iz}) E(y_i^2|x_i, \hat{\theta}) \right. \right. \]

\[ + \left. \left. \sum_{i \in A} w_i R_{iy} y_i^2 - \bar{YY}_N|F_N \right) \right] \]

\[ = E \left[ E \sum_{i \in U} (1 - R_{iy}) R_{iz} E(y_i^2|z_i, x_i, \hat{\theta}) + \sum_{i \in U} (1 - R_{iy})(1 - R_{iz}) E(y_i^2|x_i, \hat{\theta}) \right. \]

\[ - \sum_{i \in U} (1 - R_{iy}) y_i^2 \right] \]

\[ \approx E \left[ \sum_{i \in U} (1 - R_{iy}) R_{iz} E(y_i^2|z_i, x_i, \theta) + \sum_{i \in U} (1 - R_{iy})(1 - R_{iz}) E(y_i^2|x_i, \theta) \right. \]

\[ - \sum_{i \in U} (1 - R_{iy}) y_i^2 \right] \]

\[ = E \left[ \sum_{i \in U} (1 - R_{iy}) R_{iz} y_i^2 + \sum_{i \in U} (1 - R_{iy})(1 - R_{iz}) y_i^2 - \sum_{i \in U} (1 - R_{iy}) y_i^2 \right] = 0 \]

Similarly, \( E(\bar{z}z_I - \bar{ZZ}_N) \approx 0. \)
Finally,

\[
E(\overline{y}_I - \overline{Y}_N) = E \left( E(\overline{y}_I - \overline{Y}_N|F_N) \right)
\]

\[
= E \left[ \sum_{i \in A} \sum_{t \in A_{D,i}} w_i (1 - R_{iy}) R_{iz} y_t z_i + \sum_{i \in A} \sum_{t \in A_{D,i}} w_i R_{iy} (1 - R_{iz}) w_t^* z_i y_i \\
+ \sum_{i \in A} \sum_{t \in A_{D,i}} w_i (1 - R_{iy}) (1 - R_{iz}) w_t^* y_t z_i + \sum_{i \in A} w_i R_{iy} R_{iz} y_i z_i - \overline{Y}_N|F_N \right]
\]

\[
= E \left[ \sum_{i \in A} w_i (1 - R_{iy}) R_{iz} E(y_i|z_i, x_i, \hat{\theta}) z_i + \sum_{i \in A} w_i R_{iy} (1 - R_{iz}) E(z_i|y_i, x_i, \hat{\theta}) y_i \\
+ \sum_{i \in A} w_i (1 - R_{iy}) (1 - R_{iz}) E(y_i z_i|, x_i, \hat{\theta}) - \sum_{i \in A} w_i R_{iy} R_{iz} y_i z_i \right]
\]

\[
\approx E \left[ \sum_{i \in U} (1 - R_{iy}) R_{iz} E(y_i|z_i, x_i, \theta) z_i + \sum_{i \in U} R_{iy} (1 - R_{iz}) E(z_i|y_i, x_i, \theta) y_i \\
+ \sum_{i \in U} (1 - R_{iy}) (1 - R_{iz}) E(y_i z_i|, x_i, \theta) - \sum_{i \in U} (1 - R_{iy} R_{iz}) y_i z_i \right]
\]

\[
= E \left[ \sum_{i \in U} (1 - R_{iy}) R_{iz} y_i z_i + \sum_{i \in U} R_{iy} (1 - R_{iz}) y_i z_i \\
+ \sum_{i \in U} (1 - R_{iy}) (1 - R_{iz}) y_i z_i - \sum_{i \in U} (1 - R_{iy} R_{iz}) y_i z_i \right] = 0
\]

As a result, our imputed estimators are approximately unbiased estimators for the finite population parameters in (3.2).

7. Discussion of Practical Issues

In order to use the proposed fractional imputation method, one must construct weights satisfying (4.3) and (4.5). It is important to avoid the extreme weights because...
applying these weights to make estimates for various domain means and proportions may produce unrealistic estimates for some domains and proportions. In this section, we consider the method of constructing fractional weights to avoid the chance of extreme weights like negative weights.

We want to select fractional weights \( w^*_t \) satisfying (4.3) and (4.5) with \( 0 \leq w^*_t \leq 1 \) for \( t \in A_{D,i} \). This leads to a constrained minimization problem that can be solved by Lagrange multipliers. We must minimize the following expression: for each missing unit \( i \),

\[
Q(w^*_t) = d(w^*_t, \alpha_t) - \lambda' T(w^*) - \lambda_0 \left( \sum_{t \in A_{D,i}} w^*_t - 1 \right), \tag{7.1}
\]

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

Various distance measures can be used for our problem. Specifically, Hellinger distance and Entropy distance measures can be applied for nonnegative weights. A common distance measure between two sets of probabilities is Entropy measure,

\[
d(w^*_t, \alpha_t) = \sum_{t \in A_{D,i}} w^*_t \log \left( \frac{w^*_t}{\alpha_t} \right). \tag{7.2}
\]

Then we need to solve the following expression under entropy distance measure:

\[
\log(Mw^*_t) + 1 - \lambda' \frac{\partial}{\partial w^*_t} T(w^*) - \lambda_0 = 0, \tag{7.3}
\]

subject to the constraints \( T(w^*) = 0 \) and \( \sum_{t \in A_{D,i}} w^*_t = 1 \).

In the case of (4.3), implying that only the \( j^{th} \) variable is missing among \( p \) variables, \( T(w^*) \) can be written as

\[
T(w^*) = \sum_{t \in A_{D,i}} w^*_t \left( 1, y_{tj} - E(y_{tj}|\text{obs}, \hat{\theta}), y^2_{tj} - E(y^2_{tj}|\text{obs}, \hat{\theta}) \right) \tag{7.4}
\]
and the expression in (7.3) can be reduced to

$$\log(Mw^*_i) + 1 - \lambda_1 y_{ij} - \lambda_2 y^2_{ij} - \lambda_0 = 0.$$  \hfill (7.5)

Using (7.5) and \(\sum_{t \in A_{D,i}} w^*_t = 1\), the adjusted fractional weights can be written as

$$w^*_t = \frac{e^{\hat{\lambda}_1 y_{ij} + \hat{\lambda}_2 y^2_{ij}}}{\sum_{t \in A_{D,i}} e^{\hat{\lambda}_1 y_{ij} + \hat{\lambda}_2 y^2_{ij}}},$$  \hfill (7.6)

where the \(\hat{\lambda}_k\) \((k = 1, 2)\) are the solutions to \(T(w^*) = 0\) in (7.4). In general, a Newton-Raphson method can be used to solve the nonlinear equations \(T(w^*) = 0\). The resulting fractional weights will be positive with \(0 \leq w^*_t \leq 1\), \(t \in A_{D,i}\) and satisfy the constraints (4.3) or (4.3).

Note that the Euclidean distance always has a solution in (7.1) but the resulting weights can be negative and extremely large. Otherwise, the Entropy distance measure is guaranteed to obtain non-negative weights but a solution is not guaranteed for some unlucky samples of donors. To avoid “unlucky”, one approach is to use a modified sampling mechanism to select donors where the first and second moments in the donor sets are possibly close to those of the conditional distribution given by the observed data and the estimated values of parameters.

8. Variance Estimation

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

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

(8.1)

with $\xi_i$ being any component of the matrix $y_i y_i'$, $\hat{\xi}^{(k)}_n$ is the $k^{th}$ estimate of $\xi_N$, based on the observation included in the $k^{th}$ replicate, $K$ is the number of replicates and $c_k$ is a factor associated with replicate $k$ determined by the replication method. When the original estimator $\hat{\xi}_n$ is a linear estimator, the $k^{th}$ replicate estimate of $\hat{\xi}_n$ can be written as

$$
\hat{\theta}^{(k)}_n = \sum_{i \in A} w^{(k)}_{it} \xi_i,
$$

(8.2)

where $w^{(k)}_{it}$ denotes the replicate weight for the $i^{th}$ unit of the $k$ replication.

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

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

(8.3)

where

$$
\hat{\xi}_{I,n} = \sum_{i \in A} \sum_{t \in A_D} w_i w^{*}_{it} \xi_i
$$

and

$$
\hat{\xi}^{(k)}_{I,n} = \sum_{i \in A} \sum_{t \in A_D} w^{(k)}_i w^{*}_{it} \xi_i.
$$

The replicated fractional weights $w^{*}_{ij}$ in (8.3) are to be constructed using a regression weighting technique that leads to a solution satisfying the following constraints. For $R_{ij} = 0$ and $R_{ik} = 1, j \neq k,$

$$
\sum_{t \in A_D} w^{*}_{it} (1, y_{ij}, y_{ij}^2) = \left(1, E(y_{ij}|y_{obs,i}, \phi^{(k)}), E(y_{ij}^2|y_{obs,i}, \phi^{(k)})\right).
$$

(8.4)
For \( R_{ij} = 0, R_{ik} = 0 \) and \( R_{is} = 1 \) for \( s \neq j, k \),

\[
\sum_{t \in A_{D,i}} w_{it}^{(k)} (1, y_{ij}, y_{ik}, y_{ij}^2, y_{ik}^2) = (8.5)
\]

\[
\begin{align*}
&1, E(y_{ij} \mid y_{obs,i}, \hat{\theta}^{(k)}), E(y_{ik} \mid y_{obs,i}, \hat{\theta}^{(k)}), \\
&E(y_{ij} y_{ik} \mid y_{obs,i}, \hat{\theta}^{(k)}), E(y_{ij}^2 \mid y_{obs,i}, \hat{\theta}^{(k)}), E(y_{ik}^2 \mid y_{obs,i}, \hat{\theta}^{(k)})
\end{align*}
\]

where \( \hat{\theta}^{(k)} \) is the \( k^{th} \) replicate estimate of \( \hat{\theta} \) using the EM algorithm on the replicate sample.

Provided that the variance estimator of the complete estimator in (8.1) is consistent, the proposed variance estimator of the FI estimator is also consistent for the finite population means and covariances.

### 9. Simulation

In order to demonstrate the performance of the proposed estimators, we generate independently \( B = 5,000 \) random samples of size 200 from an infinite population with with three variables \( U_i = (Y_i, Z_i, X_i) \) from a trivariate normal distribution with the mean vector \( \mu = (1, 2, 3) \) and covariance matrix \( \Sigma \) with entries \( \sigma = (\sigma_{yy} = 1, \sigma_{yz} = 0.8, \sigma_{yx} = 1, \sigma_{zz} = 2, \sigma_{zx} = 1.5, \sigma_{xx} = 2) \). In addition, an indicator of membership in a domain, \( D_i \), is generated from the uniform (0, 1) distribution, independent of \( Y_i, Z_i \) and \( X_i \). The domain will be defined by \( D_i \) being below a set cutoff value. From each finite population, we also generated response indicator variables \( R_{1i} \) and \( R_{2i} \) from a Bernoulli distribution with the response rates \( p_1 = 0.65 \) and \( p_2 = 0.55 \), independently. The variable \( Y_i \) is observed if and only if \( R_{1i} = 1 \). The variable \( Z_i \) is observed if and only if \( R_{2i} = 1 \). The probability of responding to both variables is then \( 0.55 \times 0.65 = 0.3575 \), or 35.75%. In simulations, the average rate of responding to both variables was approximately 36.6%.

For the comparison, we used following methods:

**ML** Maximum Likelihood Estimation using EM algorithm.
FI  Fractional Imputation Estimation proposed in this section with $M = 10$ donors.

MI  Multiple imputation with $M = 10$ repeated imputations.

The three methods are described in more detail below.

In ML, we used estimates based on complete data set (both $Y$ and $Z$ are observed) as the starting values. Then we found the ML estimates of parameters using the EM algorithm.

In the case of FI, the selection of donors must be done carefully. The fractional weights constructed by the regression weighting technique in FI can be quite variable, producing some large weights, or even negative weights to satisfy the constraints (4.3) and (4.5). In this simulation, we used a slightly modified selection method based on the nearest neighbor criterion and simple random sampling. The nearest neighbor criterion is used for avoiding some extreme weights. Simple random sampling without replacement is used for preserving the observational distribution, instead of relying on the model to generate simulated values for imputation. In particular, for missing unit $j$, two closest donors are selected where one is the closest one to $E(U_{mis}|U_{obs}, \hat{\theta})$ among the set of observed unit having $U_{obs}$-values greater than $E(U_{mis}|U_{obs}, \hat{\theta})$ and the other one is the closest one to $E(U_{mis}|U_{obs}, \hat{\theta})$-value among the set of observed unit having $U_{obs}$-values less than $E(U_{mis}|U_{obs}, \hat{\theta})$. After selecting two donors, the $M - 2$ donors are selected with simple random sampling without replacement. When some of the final fractional weights $w_{ij}^*$ are still negative or extreme, then the algorithm for producing nonnegative fractional regression weights proposed by Paik and Larsen (2007) was applied to produce nonnegative fractional weights satisfying (4.3) and (4.5).

For MI, the missing values are generated from the posterior predictive distribution of the data given the observed values. The method of multiple imputation for the multivariate normal model is used as follows:
MI-1. For each repetition of the imputation, \( k = 1, \ldots, M \), draw

\[
\Sigma^*_{(k)} | U_{obs} \sim \text{i.i.d. \text{Inverse-Wishart}}_{v-1}(S)
\]

where \( v \) is the size of the set \( A_{rr} \) and \( S \) is the sum of squares matrix about the sample mean on complete data \( A_{rr} \):

\[
S = \sum_{A_{rr}} (U_i - \bar{U}_r)(U_i - \bar{U}_r)',
\]

where \( \bar{U}_r \) is the mean of \( U_i \) on \( A_{rr} \).

MI-2. Generate

\[
\bar{U}^*_{(k)} | (U_{obs}, \Sigma^*_{(k)}) \sim \text{i.i.d. } N(\bar{U}_r, \Sigma^*_{(k)}).
\]

MI-3. For missing unit \( j \),

\[
e^*_{j(k)} | (\bar{U}^*_{(k)}, \Sigma^*_{(k)}) \sim \text{i.i.d. } N(0, \Sigma^*_{(k)}).
\]

Then \( U_{j(k)} = E(U_{j} | U_{obs}, \bar{U}^*_{(k)}) + e^*_{j(k)} \) is the value associated with unit \( j \) for \( k^{th} \) imputation.

MI-4. Repeat steps 1-3 independently \( M \) times.

10. Simulation Results

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

1. \( \bar{Y}_N, \bar{Z}_N, S_{yy,N}, S_{zz,N}, S_{yz,N}, S_{yx,N} \) and \( S_{zx,N} \),

2. \( \bar{Y}_{D,N} \) and \( \bar{Z}_{D,N} \) are means of \( Y \) and \( Z \) where \( D < 0.45 \),

3. \( P_{y,N} \) = proportion of \( Y > 1.65 \), and

4. \( P_{z,N} \) = proportion of \( Z < 1.38 \).
For variance estimation, we have considered the FI estimator and the MI estimator of variance. For the FI variance estimator, we used the jackknife variance estimation method discussed in previous section. In case of the MI variance estimator, the simple variance formula of Rubin is used.

The Monte Carlo results for 5,000 samples generated are given Table B.1 and Table B.2. Table B.1 shows the mean and variance of the point estimators for three methods. The properties of the variance estimators (MI and FI) are given in Table B.2. Table B.2 shows the relative bias and t-statistic for the variance estimators. The relative bias of the variance estimate is estimated by

$$RB(\%) = \frac{E_B(\hat{V}) - V_B(\hat{\theta}_I)}{V_B(\hat{\theta}_I)} \times 100,$$

where $E_B$ and $V_B$ are the Monte Carlo expectation and variance. The t-statistics for testing $H_0: E(\hat{V}) = Var(\hat{\theta}_I)$ is

$$t - statistic = \sqrt{B} \frac{E_B(\hat{V}) - V_B(\hat{\theta}_I)}{\sqrt{V_B(\hat{V} - (\hat{\theta}_I - E_B(\hat{\theta}_I))^2)}}$$

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

In Table B.1, the proposed FI estimator shows more efficacy than the MI estimator for all parameters except the domain means and proportions. For improving the efficiency of the estimation of these parameters, the imputed values may be generated from the assumed conditional distribution given the observed data.

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

11. Conclusion

Based on the simulation results, the proposed fractional imputation method seems to be a good imputation method because it retains the desirable properties of maximum likelihood estimation when estimating the parameters of the superpopulation model, uses actually observed values, and produces a single set of general purpose replicate fractional weights. In addition, it provides reasonable estimates for other parameters that were not included in the imputation models. As with other fractional imputation methods, an easy-to-use data set was constructed for general purpose estimation. For the completed data set constructed by the proposed procedure, the standard estimates at the aggregate level of analysis are equivalent to model-based imputation estimates based on maximum likelihood for parameters in the imputation model.
Table B.1 Monte Carlo means and variances for imputation estimators, based on 5,000 samples.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Imputation Method</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_N$</td>
<td>Complete Sample</td>
<td>1.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>1.02</td>
<td>0.0062</td>
</tr>
<tr>
<td></td>
<td>FI(M=10)</td>
<td>1.02</td>
<td>0.0062</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>1.02</td>
<td>0.0065</td>
</tr>
<tr>
<td>$Z_N$</td>
<td>Complete Sample</td>
<td>2.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>2.03</td>
<td>0.0137</td>
</tr>
<tr>
<td></td>
<td>FI(M=10)</td>
<td>2.03</td>
<td>0.0137</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>2.03</td>
<td>0.0147</td>
</tr>
<tr>
<td>$S_{yy,N}$</td>
<td>Complete Sample</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>1.00</td>
<td>0.0146</td>
</tr>
<tr>
<td></td>
<td>FI(M=10)</td>
<td>1.00</td>
<td>0.0146</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>0.97</td>
<td>0.0153</td>
</tr>
<tr>
<td>$S_{zz,N}$</td>
<td>Complete Sample</td>
<td>2.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>2.02</td>
<td>0.0063</td>
</tr>
<tr>
<td></td>
<td>FI(M=10)</td>
<td>2.02</td>
<td>0.0063</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>2.04</td>
<td>0.0087</td>
</tr>
<tr>
<td>$S_{yz,N}$</td>
<td>Complete Sample</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>0.82</td>
<td>0.0207</td>
</tr>
<tr>
<td></td>
<td>FI(M=10)</td>
<td>0.82</td>
<td>0.0207</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>0.82</td>
<td>0.0213</td>
</tr>
<tr>
<td>$S_{yx,N}$</td>
<td>Complete Sample</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>1.01</td>
<td>0.0172</td>
</tr>
<tr>
<td></td>
<td>FI(M=10)</td>
<td>1.01</td>
<td>0.0172</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>1.01</td>
<td>0.0177</td>
</tr>
<tr>
<td>$S_{zx,N}$</td>
<td>Complete Sample</td>
<td>1.51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>1.51</td>
<td>0.0384</td>
</tr>
<tr>
<td></td>
<td>FI(M=10)</td>
<td>1.51</td>
<td>0.0384</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>1.51</td>
<td>0.0404</td>
</tr>
<tr>
<td>$Y_{D,N}$</td>
<td>Complete Sample</td>
<td>1.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FI(M=10)</td>
<td>1.02</td>
<td>0.0211</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>1.02</td>
<td>0.0191</td>
</tr>
<tr>
<td>$Z_{D,N}$</td>
<td>Complete Sample</td>
<td>2.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FI(M=10)</td>
<td>2.03</td>
<td>0.0346</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>2.03</td>
<td>0.0245</td>
</tr>
<tr>
<td>$P_{y,N}$</td>
<td>Complete Sample</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FI(M=10)</td>
<td>0.26</td>
<td>0.0012</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>0.26</td>
<td>0.0012</td>
</tr>
<tr>
<td>$P_{z,N}$</td>
<td>Complete Sample</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FI(M=10)</td>
<td>0.32</td>
<td>0.0013</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>0.32</td>
<td>0.0013</td>
</tr>
</tbody>
</table>
Table B.2  Relative biases and t-statistics for the variance estimators, based on 5,000 samples.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Imputation Method</th>
<th>RB(%)</th>
<th>t-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_N$</td>
<td>FI(M=10)</td>
<td>0.51</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>-1.61</td>
<td>-0.80</td>
</tr>
<tr>
<td>$Z_N$</td>
<td>FI(M=10)</td>
<td>4.06</td>
<td>2.05</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>5.09</td>
<td>2.47</td>
</tr>
<tr>
<td>$S_{yy,N}$</td>
<td>FI(M=10)</td>
<td>-2.66</td>
<td>-1.23</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>-2.76</td>
<td>-1.38</td>
</tr>
<tr>
<td>$S_{zz,N}$</td>
<td>FI(M=10)</td>
<td>2.34</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>-3.07</td>
<td>1.48</td>
</tr>
<tr>
<td>$S_{yz,N}$</td>
<td>FI(M=10)</td>
<td>1.07</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>5.07</td>
<td>2.50</td>
</tr>
<tr>
<td>$S_{yx,N}$</td>
<td>FI(M=10)</td>
<td>-1.28</td>
<td>-0.60</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>3.92</td>
<td>1.91</td>
</tr>
<tr>
<td>$S_{zx,N}$</td>
<td>FI(M=10)</td>
<td>3.51</td>
<td>1.22</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>8.47</td>
<td>4.25</td>
</tr>
<tr>
<td>$Y_{D,N}$</td>
<td>FI(M=10)</td>
<td>-6.12</td>
<td>-3.13</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>14.68</td>
<td>7.29</td>
</tr>
<tr>
<td>$Z_{D,N}$</td>
<td>FI(M=10)</td>
<td>-7.21</td>
<td>-3.97</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>27.93</td>
<td>13.28</td>
</tr>
<tr>
<td>$P_{y,N}$</td>
<td>FI(M=10)</td>
<td>-1.51</td>
<td>-0.71</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>14.35</td>
<td>6.49</td>
</tr>
<tr>
<td>$P_{z,N}$</td>
<td>FI(M=10)</td>
<td>1.60</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>MI(M=10)</td>
<td>24.00</td>
<td>12.02</td>
</tr>
</tbody>
</table>
WEIGHT ADJUSTMENTS FOR FRACTIONAL REGRESSION HOT DECK IMPUTATION

A paper appeared as a proceedings paper of the 2007 Joint Statistical Meetings.

Minhui Paik and Michael D. Larsen

Abstract

Fractional regression hot deck imputation (FRHDI), suggested by J. K. Kim, imputes multiple values for each instance of a missing dependent variable. The imputed values are equal to the predicted value based on the fully observed cases plus multiple random residuals chosen from the set of empirical residuals. Fractional weights are chosen to enable variance estimation and to preserve the correlation among independent and dependent variables. The FRHDI method can be viewed as a special case of fractional hot deck imputation (FHD). In some circumstances with some starting weight values, existing procedures for computing FRHDI weights can produce negative values. We discuss procedures for constructing nonnegative adjusted fractional weights for FRHDI.

1. Introduction

Consider a population of $N$ elements identified by a set of indices $U = \{1, 2, \ldots, N\}$. Associated with unit $i$ of the population are two study variables, $y_i$ and $x_i$, where every
\( x_i \) is complete and some \( y_i \) are missing. Let \( A \) denote the set of indices of the elements in a sample selected by the chosen sampling mechanism. Responses \( y_i \) are obtained from the selected sample according to the response mechanism. Let the population quantity of interest be \( \theta_N = \theta(y_1, y_2, \ldots, y_N) \) or \( \theta_N = \theta((y_1, x_1), (y_2, x_2), \ldots, (y_N, x_N)) \). Under complete response, an unbiased linear estimator of \( \theta_1 = N^{-1} \sum_{i=1}^{N} y_i \) is

\[
\hat{\theta}_1 = \sum_{i \in A} w_i y_i \tag{1.1}
\]

where \( w_i \) is a sampling weight for unit \( i \) that depends on the sampling mechanism.

Another parameter of interest is \( \theta_2 = N^{-1} \sum_{i=1}^{N} y_i x_i \). An unbiased linear estimator of \( \theta_2 \) is

\[
\hat{\theta}_2 = \sum_{i \in A} w_i y_i x_i \tag{1.2}
\]

Hot deck imputation assigns the value of \( y_i \) for respondents to missing \( y \)-values for nonrespondents. One of the main considerations for hot deck imputation is how best to select the donor values. Many hot deck imputation procedures select donor values at random in the same imputation cell, which can be constructed by partitioning the sample using auxiliary variables known for both the respondents and the nonrespondents. An advantage of this method is that the actual observed values are used for imputation and, assuming some homogeneity within cells, imputations are realistic. The performance of hot deck imputation depends on the quality of available donors for the missing cases.

The method of stochastic regression imputation replaces a missing value by a predicted value plus a residual, which is drawn to reflect uncertainty in the predicted value. In notation, let \( A_r \subseteq A \) be the set of indices for respondents and \( A_M \subseteq A \) contain the indices for missing values. The imputed value \( y_i^* \), \( i \in A_M \), is

\[
y_i^* = \hat{y}_i + \hat{e}_j^* \tag{1.3}
\]

where \( \hat{y}_i \) is the predicted value of \( y_i \) and \( \hat{e}_j^* \) is an imputed residual randomly selected from \( \{ \hat{e}_j^* = y_j - \hat{y}_j, j \in A_r \} \). The predictions, \( \hat{y}_i \) for \( i \in A_M \) and \( \hat{y}_j \) for \( j \in A_r \), are based
on the relationship between $x$ and $y$ for cases in $A_r$.

Stochastic regression imputation maintains the distribution of the variables in the sense of maintaining the observed relationship between $y$ and $x$ and allows for the estimation of distributional quantities. However, this method is potentially more sensitive to model violations than methods based on implicit models, such as hot deck imputation. In addition, the imputed value is not necessarily one of the actually occurring values, which in some situations can be seen as a negative.

There is one further disadvantage of imputing a single value for each missing value. Single imputation cannot represent uncertainty due to imputation. Multiple imputation methods, including multiple imputation (Rubin 1978, 1987, 1996) and fractional imputation (Kim and Fuller 2004), consider multiple possible values for each missing value. The variability in imputed values is used to in effect quantify uncertainty due to imputation. Imputation procedures also vary in terms of the amount of variability introduced through the process of producing imputations.

Brick and Kalton (1996) studied two methods for reducing the imputation variance which comes from the random component of the variance of the estimator arising from imputation. One method is implemented through the sample design used for selecting donors within each imputation cell. For example, selection without replacement is less variable than selection with replacement. The other method is to use fractional imputation (Kalton and Kish, 1984; Fay 1996), which uses more than one donor for a recipient and assign fractional survey weights to multiple donors. Fractional imputation was suggested as a method for expressing uncertainty due to imputation and reducing imputation variance. It later provides as useful tool for variance estimation by Kim and Fuller (2004). However, fractional hot deck imputation can not preserve the correlation structure among two or more quantitative variables except for the variables that define imputation cells. As a result, a relationship between an independent and dependent variable could be weakened due to simple hot deck imputation, even if the imputations
are done multiple times.

Kim (2007) suggested Fractional Regression Hot Deck Imputation (FRHDI) in order to combine the advantages of hot deck imputation and regression imputation within the framework of fractional imputation. The procedure for combining the two imputation methods takes the form of fractional hot deck imputation with a suitable choice of fractional weights. Consequently, FRHDI preserves the correlation structure and uses observed values for imputation. In addition, a jackknife variance estimation technique developed by Kim and Fuller (2004) can be applied for variance estimation.

It is known, however, that the weights constructed by the regression weighting method can be vary, producing some large weights, or can be negative. A large weight on donors can result in large imputation variance for some estimates. In particular, estimates within a domain can be highly variable if some weights are extreme. A negative fractional weight can be seriously problematic for estimating the variance of imputed estimators.

In this paper, we modify an iterative regression procedure suggested by Huang and Fuller (1978) to construct nonnegative fractional weights and to place bounds on the fractional weights. The review of FRHDI is described in Section 2. The proposed method of constructing nonnegative fractional weights is discussed in Section 3. Simulation results are reported in Section 4. Section 5 is a discussion and summary.

2. Fractional Regression Hot Deck Imputation

One can indicate the donors for missing value $y_j$, $j \in A_m$ through indicator variables $d = \{d_{ij}; i \in A_r\}$. Let the indicator variable $d_{ij}$ take the value one if $y_i$ is used as a donor for the missing $y_j$ and take the value zero otherwise. The sampling weight $w_j$ is distributed to the donors with $d_{ij} = 1$. Let $w_{ij}^*$ be the fractional weight allocated to donor $i$ for recipient $j$. The sum of fractional weights for each missing value is required
to be one. Assume that the finite population \( U \) is made up of \( G \) imputation cells and the cell regression model is appropriate for each cell. That is, for \( i \in A_g \), the \( g \)th imputation cell,

\[
E(y_i | x_i) = \beta_{0g} + \beta_1 x_i \quad \text{and} \quad V(y_i | x_i) = \sigma^2_g. \tag{2.1}
\]

To apply regression imputation to fractional imputation, the weighted mean of the imputed values using stochastic regression imputation is used to impute the missing data. Let missing values and observed values in cell \( g \) be indicated by \( A_mg \) and \( A_rg \), respectively. Let \( A_Dgj \) is the set of indices of imputed values for \( j \in A_mg \) where \( A_Dgj = \{ i; d_{ij} = 1 \} \).

For \( j \in A_mg \), the imputed value for missing \( y_j \) is

\[
y^*_j = \sum_{i \in A_Dgj} w^*_ij \left( \hat{y}_j + \hat{e}_i \right). \tag{2.2}
\]

In the above formula, \( \sum_{i \in A_rg} w^*_ij d_{ij} = \sum_{i \in A_Dgj} w^*_ij \).

The imputed estimator of \( \theta_1 \) under fractional imputation can be constructed as follows:

\[
\hat{\theta}_{I1} = \sum_{g=1}^{G} \left[ \sum_{i \in A_rg} w_i y_i + \sum_{j \in A_mg} w_j y^*_j \right] \\
= \sum_{g=1}^{G} \left[ \sum_{i \in A_rg} w_i y_i + \sum_{j \in A_mg} w_j \sum_{i \in A_Dgj} w^*_ij \left( \hat{y}_j + \hat{e}_i \right) \right] \\
= \sum_{g=1}^{G} \left[ \sum_{i \in A_rg} w_i y_i + \sum_{j \in A_mg} \sum_{i \in A_Dgj} w_j w^*_ij \left( \hat{y}_j + \hat{e}_i \right) \right].
\]

When \( \hat{y}_j \) is linear function of \( x_j \), then the weighted mean of the imputed values can be written as

\[
\sum_{i \in A_Dgj} w^*_ij \left( \hat{y}_j + \hat{e}_i \right) = \sum_{i \in A_Dgj} w^*_ij \left( y_i + (x_j - x_i) \hat{\beta} \right) \\
= \sum_{i \in A_Dgj} w^*_ij y_i + \hat{\beta} \sum_{i \in A_Dgj} w^*_ij (x_j - x_i). \tag{2.3}
\]
The second term on the right side of (2.3) can be zero by using adjusted fractional weights. The sufficient and necessary condition for taking the form of the fractional Hot deck imputation is

$$\sum_{i \in A_{Dgj}} w_{ij}^*(1, x_i) = (1, x_j) \quad \text{for } j \in A_{mg} \quad (2.4)$$

The method of constructing fractional weights with constrain (2.4) will be presented in section (3). The proposed procedure by Kim (2007) can be motivated as follows. First, all missing values can be imputed by observed values like hot deck imputation because of adjusting the fractional weight associated with donors in observed sample instead of imputing unobserved values. Second, it is easy to estimate the variance of the imputed estimator by applying a consistent replication variance estimation procedure with fractional imputation suggested by Kim and Fuller (2004). Third, FRHDI is expected to be more robust against model violations in that FRHDI use fractional weights constructed by the relationship between the auxiliary variable $x_i$ in $A_r$ and $x_j$ in $A_m$ not depend on the estimator of $\beta$ based on the explicit models.

The FRHDI estimator can be expressed in the form of the fractional hot deck imputed estimator as follows:

$$\hat{\theta}_{I1} = \sum_{g=1}^{G} \left( \sum_{i \in A_{r,g}} w_{ij} y_i + \sum_{j \in A_{mg}} \sum_{i \in A_{Dgj}} w_{ij}^* y_i \right)$$

where the $w_{ij}$ satisfy condition (2.4).

Under the cell regression model (2.1) and ignorable response mechanism, the conditional expectation of the FRHDI estimator of $\theta_1$ is

$$E(\hat{\theta}_{I1} - \hat{\theta}_1 | A, A_r, \chi) = E \left( \sum_{g=1}^{G} \sum_{j \in A_{mg}} w_j \left[ \left( \sum_{i \in D_{gj}} w_{ij}^* y_i \right) - y_j \right] | A, A_R, \chi \right)$$

$$= \sum_{g=1}^{G} \sum_{j \in A_{mg}} w_j \left[ \sum_{i \in D_{gj}} w_{ij}^* (\beta_0 g + \beta_1 x_i) - \beta_0 g + \beta_1 x_j \right]$$
which is equal to zero by (2.4), where $\chi = \{(1, x_i : i \in A)\} \text{ and } D_{gj} \subseteq A_{rg}$. Therefore, the FRHDI estimator is unbiased provided that the complete estimator $\hat{\theta}_n$ is unbiased. Similar algebra can be used to write the imputed estimator of $\theta_2$ as a fractional hot deck (FHD) estimator and the conditional unbiasedness also holds for $\theta_{12}$. For variance estimation of FRHDI estimator, Kim (2007) modified a replication variance estimator by Kim and Fuller (2004) that changes the fractional replicate weights so that the expected value of the sum of square is changed by the proper amount replacing the cell mean model by the cell regression model.

3. Construction of Regression Fractional Weights

In order to use fractional regression hot deck imputation, one must construct weights $w_{ij}^*, i \in A_{D_{gj}}$ for each $j \in A_{mg}$ such that $0 < w_{ij}^* < 1$, $\sum_{i \in A_{D_{gj}}} w_{ij}^* = 1$ and formula (2.4) holds. It has been noted that numerical procedures for computing weights sometimes encounter problems. It is possible that suitable weights might not exist. It also is a fact that numerical procedures that do not directly incorporate all the constraints, can produce weights that are negative, which is undesirable. Section 3.1 discusses weight computation. Section 3.2 presents a modification to methods when standard computational methods encounter a problem.

3.1. Introduction to Regression Weighting

Kim (2007) suggested the regression weighting method to construct the fractional weights satisfying the constraint (2.4). This method can be viewed as a calibration technique. This procedure for constructing the fractional weights is to minimize a function of the distance between an initial weight $\alpha_{ij}$ and a final fractional weight $w_{ij}^*$ subject to the constraint (2.4). Let $\alpha_{ij}$ be any initial fractional weights satisfying $\sum_{i \in A_{D_{gj}}} \alpha_{ij} = 1$. A common choice is $\alpha_{ij} = 1/M$ for $j \in A_m$ where $M$ is the number
of donors used for fractional imputation. Let the distance function between $\alpha_{ij}$ and $w_{ij}^*$ be $Q(\alpha_{ij}, w_{ij}^*) = \sum_{i \in A_{Dg}} \alpha_{ij}^{-1}(\alpha_{ij} - w_{ij}^*)^2$. Then the problem is to minimize

$$Q(\alpha_{ij}, w_{ij}^*)$$

subject to the constraints

$$\sum_{i \in A_{Dg}} w_{ij}^*(1, x_i) = (1, x_j)$$

and

$$0 < w_{ij}^* < 1, j \in A_{mg}.$$ (3.1)

By using the Lagrange multiplier method, the solution of (8) is

$$w_{ij}^* = \alpha_{ij} + (x_j - \bar{x}_{Ij})S_{xx,j}^{-1} \alpha_{ij}(x_i - \bar{x}_{Ij})$$ (3.2)

where

$$S_{xx,j} = \sum_{i \in A_{Dg}} \alpha_{ij}(x_i - \bar{x}_{Ij})^2$$

and

$$\bar{x}_{Ij} = \sum_{i \in A_{Dg}} \alpha_{ij}x_i.$$  

Under the calibration property $\sum_{i \in A} w_i x_i = \bar{x}_N$, not the full condition (2.4), there are several ways to construct regression weights with a reduced range of values. Huang and Fuller (1978) defined a procedure to modify the $w_i$ so that there are no negative weights and no large weights. Husain (1969) suggested quadratic programming as a procedure to place bounds on the weights. Deville and Särndal (1992) considered some objective functions (e.g., $Q$) that can be used to produce positive weights with a certain range. Park and Fuller (2005) suggested that nonnegative regression weights can be computed by a calibration technique using an initial weight, the inverse of the approximate conditional inclusion probability.

Another modification to regression weights is to relax the calibration property. This approach was studied by several authors (Husain 1969, Bardsley and Chambers 1984,
and Rao and Singh 1997). However, the constraint (2.4) is important so that it cannot be relaxed in our situation.

In this paper, we modify the method of constructing nonnegative weights with the constraints (2.4), not the calibration property, to get adjusted fractional weights. Wayne Fuller, in personal communication, has pointed out that there is no guarantee that a solution exists for the weights constructed by a quadratic programming problem with bounds on the weights. To ensure the existence of a solution, we assume that there exists at least one donor with an $x$-value greater than the value $x_j$ and one donor with an $x$-value less than the $x$-value, $x_j$, for the case with the missing $y$-value.

### 3.2. Computer Algorithm for Regression Fractional Weights

The algorithm by Huang and Fuller (1978) produces weights that are a smooth, continuous, monotone increasing function of the original least squares regression weights based upon the idea of generalized least squares. We modify their algorithm to apply for our problem. This procedure is iterative, checking the weight at each step against a user supplied criterion. The fractional weight (3.2) can be rearranged by

$$w^*_{ij} = \alpha_{ij}(1 + \phi_i) \quad (3.3)$$

where

$$\phi_i = \frac{(x_j - \bar{x}_I)S_{xx,j}^{-1}(x_i - \bar{x}_I)}{(x_j - \bar{x}_I)^2 S_{xx,j}^{-1} + (x_i - \bar{x}_I)^2}$$

An alternative computational form equivalent to weights (3.3) can be constructed by

$$w^*_{ij} = \alpha_{ij} \left(1 + \frac{z_{ij}}{1 + \varepsilon_j}\right) \quad (3.4)$$

where
\[ z_{ij} = \left( x_j - \bar{x}_{Ij} \right) \left( \sum_{i \in A_R} \alpha_{ij} (x_i - x_j)^2 \right)^{-1} (x_i - x_j) \]

\[ \bar{z}_j = \sum_{i \in A_{Dsj}} \alpha_{ij} z_{ij} \]

The regression fractional weights defined by (3.4) will be nonnegative if \( 1 + z_{ij} \geq 0 \) for every \( i \). Our computer algorithm create nonnegative fractional weights by modifying the \( z_{ij} \) such that \( |z_{ij}| < 1 \) so that \( w^*_{ij} > 0 \) and there is an positive \( L \leq 1 \) such that \( \max_{i \in A_{Dsj}} (w^*_{ij}) < L \). If the first-step fractional weights fall outside the desired range \( |z_{ij}| < 1 \), then a relatively small adjustment values are assigned to the fractional weights of donors where the auxiliary variable of donor \( x_i \) is far from that of recipient \( x_j \) and relatively large adjustment values are assigned to the fractional weights of donors where the auxiliary variable of donor \( x_i \) is close to that of recipient \( x_j \). The initial weights \( \alpha_{ij} \) are used as first-step weights. Adjustment values can be obtained by \( \gamma_i \), a "bell" shaped function (in a suitable metric) between the auxiliary variable of the donor and recipient for each donor.

The algorithm for computing the regression fractional weights is composed of the following steps. For simplicity of notation, assume there is only one imputation cell and \( G = 1 \). If there are multiple imputation cells, then implement the algorithm separately within each cell.

**STEP 1:** Calculate

\[ z_{ij} = \left( x_j - \bar{x}_{Ij} \right) \left( \sum_{i \in A_R} \alpha_{ij} (x_i - x_j)^2 \right)^{-1} (x_i - x_j) \]

and

\[ \bar{z}_j = \sum_{i \in A_{Dsj}} \alpha_{ij} z_{ij} \]

If \( |z_{ij}| < 1 \) for every \( i \in A_{Dsj} \), then the initial weights \( \alpha_{ij} \) satisfy the constraints. If not, set \( k = 1 \) and go to the next step.
STEP 2: Compute the adjusted weight for each distance $d_i$, where

$$d^{(k)}_i = \frac{4}{3} |z_{ij}|,$$

and

$$\gamma^{(k)}_i = \begin{cases} 
1 & 0 \leq d^{(k)}_i < \frac{1}{2} \\
1 - \frac{4}{5}(d^{(k)}_i - \frac{1}{2})^2 & \frac{1}{2} \leq d^{(k)}_i \leq 1 \\
\frac{4}{5}(d^{(k)}_i)^{-1} & d^{(k)}_i > 1
\end{cases}$$

The constants $4/3$ and $4/5$ are to speed convergence of the algorithm. Alternative $d$ and $\gamma$ function can be constructed.

STEP 3: Compute the new regression fractional weights

$$\lambda^{(k)}_i = \prod_{j=1}^{k} \gamma^{(j)}_i$$

$$z^{(k)}_{ij} = (x_j - \bar{x}_{ij}) \left( \sum_{i \in A_{Dj}} \alpha_{ij} \lambda^{(k)}_i (x_i - x_j)^2 \right)^{-1} \lambda^{(k)}_i (x_i - x_j)$$

$$\bar{z}^{(k)}_j = \sum_{i \in A_{Dj}} \alpha_{ij} z^{(k)}_{ij}$$

STEP 4: If $|z^{(k)}_{ij}| < 1$ for every $i \in A_{Dj}$, then

$$w^{*(k)}_{ij} = \alpha_{ij} \left( \frac{1 + z^{(k)}_{ij}}{1 + \bar{z}^{(k)}_j} \right).$$

If not, set $k = k + 1$ and go to STEP 2. Note that $z_{ij}$ is replaced by $z^{(k)}_{ij}$ in STEP 2.

The final fractional weights $w^{*(k)}_{ij}$ have the following properties. For each $j \in A_{mg}$,

(i) $0 < w^{*(k)}_{ij} < 1$ for $i \in A_{Dgj}$ and

(ii) $\sum_{i \in A_{Dgj}} w^{*(k)}_{ij}(1, x_i) = (1, x_j)$ for $j \in A_{mg}$. 
The proposed computer algorithm will produce positive fractional weights with the constraints (2.4) under the assumption that there exists at least one donor with an $x$-value greater than the value $x_j$ and one donor with an $x$-value less than the $x$-value, $x_j$, for the case with the missing $y$-value. The selection of donors have to be careful to meet the assumption for the required restrictions under our algorithm.

4. Simulation Study

This section presents the main results from two limited simulation studies. To show the performance of our procedure, we compared three imputation methods:

- FH− : FRHDI using the regression fractional weight,
- FH+ : FRHDI using the nonnegative fractional weight, and
- MI : Multiple imputation.

For fractional imputation, after computing the distance from all respondents to the unit with missing item, the different $M$ donors are selected in which the units having the smallest $M$ distance. Fractional weights $w_{ij}^*$ on FH− are calculated by the regression method (9) setting the initial fractional weight equal to $\alpha_{ij} = 1/M$. Nonnegative fractional weights are obtained by the computer algorithm in Section 3.2. For multiple imputation, $M$ repeated imputation values are drawn from the posterior predictive distribution of the missing values under a simple linear regression model given a constant prior. In this simulation, we used $M = 3, 5$ and 10.

In the first simulation, three independent variables were generated: $x_i$ from a normal distribution with $N(0, 1)$, $e_i \sim N(0, 1)$, and $z_i$ from the uniform (0,1) distribution. The dependent variable is $y_i = 2 + x_i + e_i$. We also generated a response indicator $R_i$ from a Bernoulli distribution with the response rate $p = 0.65$. The $y_i$ is observed if and only if $R_i = 1$. The $x_i$ and $z_i$ are observed throughout the sample. We used $B = 5000$
samples of size $n = 100$ to simulate properties of the procedures. Three parameters are estimated. The parameters are

$$\theta_1 = \text{mean of } Y,$$

$$\theta_2 = \text{mean of } Y \text{ where } z < 0.25,$$

$$\theta_3 = \text{proportion of } Y \leq 1\text{ and}$$

$$\theta_4 = \text{slope of } Y \text{ on } X.$$

For fractional imputation, the variance estimation method proposed by Kim (2007) was applied. Kim (2007) considers the adjusted jackknife replicates constructed by decreasing the appropriate amount of fractional weights of the imputed values associated with a deleted respondent and increasing the appropriate amount of fractional weights of the other donors we mentioned. There are certain situations where it is not possible to find the appropriate amount when using negative fractional weights. In certain cases in which there were difficulties, we used the approximated amount of adjustment for the variance estimator under FH−. The variance estimator for multiple imputation was given in Rubin (1978, 1987).

Table C.1 shows the mean and variance of the imputed estimator calculated based on the Monte Carlo sample generated by the linear regression model. The three imputation methods are unbiased for all parameters. Fractional imputation using nonnegative fractional weights and Multiple imputation is slightly more efficient than Fractional imputation using regression fractional weights. Specially, the FH− estimators for $M = 3$ have a large variance because there are no bounds for the fractional weights so that the fractional weights adjusted by regression weighting method are large and negative for some donors under the small number of donors. It imply that the large and negative fractional weights can result in large variances for point estimates. In addition, the large
and negative fractional weights can be relaxed by increasing $M$ so that the efficiency of point estimates can be improved.

In Table C.2, relative biases, t-statistics and relative variance for the variance estimators are presented. The relative bias of the variance estimator is the Monte Carlo bias (the mean of the variance estimates minus the variance of the estimates) divided by the Monte Carlo mean of the variances. The t-statistic for testing the hypothesis of zero bias is the Monte Carlo estimated bias divided by the Monte Carlo standard error of the estimated bias (the square root of the variance of the estimated biases). The relative variance is the Monte Carlo variance of the variance estimator divided by the square of the variance, where the variance is given in Table C.1. The fractional imputation variance estimation and multiple imputation variance procedures appear to be nearly unbiased for the variance of the mean. The relative variance of the MI variance estimator is nearly twice that of the FI variance estimator with nonnegative fractional weights. The FH− for $M = 3$ has a really large variance because of using the large and negative weights. This result support our proposed method in that negative fractional weights can lead serious problem for estimating the variance of imputed estimator. In Table C.2, the same patterns were observed for other parameters also. In addition, the large variance of multiple imputation estimator can be explained by the fact that the variance due to missing observation is estimated with $M − 1$ degrees-of-freedom for each $M$ imputation.

We have reasonably small relative biases for the variances of the imputed estimators $\hat{\theta}_1$ and $\hat{\theta}_2$. The fractionally imputed variance estimator is biased for the variance of the estimator of domain mean $\hat{\theta}_3$. The main source of this bias is the bias in the jackknife variance estimator for a ratio.

Table C.2 further illustrates that multiple imputation produces a seriously biased estimator of the variance of the estimator of $\theta_3$. This bias in the multiple imputation variance estimator for a domain mean was pointed out by Fay (1992). Kim and Fuller
(2004) point out that the bias in the multiple imputation variance estimator for the mean can be reduced by increasing $M$ and sample size $n$. Increasing $M$ or $n$, however, reduces only some part of the bias of the multiple imputation variance estimator for the domain mean since the MI variance estimator does not reflect the fact that the imputed values used in the domain come from observations outside the domain. Of course, if the MI procedure used the domains explicitly one could expect different results. Though the FI variance estimator for the domain mean also has a positive bias, the bias of FI is much smaller than that of MI and this bias can be reduced by increasing $M$. However, the bias of MI has little relationship to $M$.

Since the fractional weights are slightly smoothed by the regression procedure, fractional variance estimators for the variance of proportion are slightly negatively biased. MI has a large variance for variance estimator and the bias of MI cannot be reduced by increasing $M$.

One additional result can be mentioned. The variance estimators for FH+ are more stable than those for FH− and MI. FH+ has the relatively small variance of the variance estimators on all four parameters. Since the proposed method improve the performance of FRHDI under the small number of donor size, our proposal is useful because the donor size is generally limited in practice.

In the second simulation study, the samples are generated from the quadratic regression model $y_i = 2 + \sqrt{0.5}(x_i^2 - 1) + e_i$ where $x_i$ and $e_i$ are as before. The same $z_i$ and $R_i$ variables generated in the first simulation were used. We expect the FRHDI be quite robust against the misspecification of the imputation model since the imputed values for fractional imputation are selected from nearest neighborhoods and thus correspond to a local (rather than global) linear model. To demonstrate the robustness of FRHDI with nonnegative fractional weights, we used the simple linear regression model not the true quadratic model for imputation. Of course, if one used the quadratic model for imputation, then results would be different. The parameters we consider in simulation
2 are the same as in simulation 1.

Table C.3 shows the performance of the point estimator under simulation 2. The Monte Carlo results are in general agreement with our expectation. Fractional imputation methods with nonnegative fractional weights are approximately unbiased and have more efficiency than that of FH− and MI for all parameters except slope, specially, for $M = 3$. Fractional imputation methods are more robust against the failure of the imputation model than multiple imputation. Multiple imputation estimators show big biases for all parameters since the assumption required for MI estimation procedure are not satisfied under simulation 2. This result indicates that, as expected, MI is very sensitive to the model used for imputation. Bias of multiple imputation estimator could be improved some of degree if the imputed values were generated by the local simple linear regression model. The robustness of FRHDI against the failure of the imputation model can be explained by that the imputed values for fractional imputation are selected from the observed unit with $M$ closest $x$-values.

In Table C.4 the biases and t-statistics of three imputation methods are illustrated under simulation 2. Under misspecification of imputation model, the proposed method still improve the performance of FRHDI based on three statistics in Table C.4. This improvement is increased as decreasing $M$. The FH+ variance estimator shows much smaller relative biases, t-statistic and relative variances than that of Multiple imputation.

5. Summary and Discussion

We have discussed a procedure for constructing nonnegative adjusted fractional weights for fractional regression hot deck imputation (FRHDI). Fractional imputation permits the adjustment of fractional weights to incorporate the auxiliary variable information into the imputed estimator. Kim (2007) suggested the method of adjusting fractional weight to reflect the information of auxiliary variable using regression weighting
method. However, it is known that the weights constructed by the regression weighting method can be large weight or negative. Under fractional imputation, large or negative fractional weights can be seriously problematic for estimating some parameter and the variance of imputed estimators. Specially, this problem is more serious when the number of donor size is small. FRHDI with regression fractional weights may not be preferred since the donor size is generally limited in practice. In this paper, we suggest an iterative regression procedure for nonnegative fractional weights. In the limited simulation, the proposed method performs better than FRHDI with regression fractional weights and Multiple Imputation. To apply the suggested method for FRHDI, the selection of donors have to be careful to meet the assumption of the required restrictions under the proposed method and to keep the distribution of observed units since the weights of selecting unit as a donor are slightly smoothed by the regression procedure.

FRHDI with the nonnegative fractional weights is more efficient for some parameters and gives variance estimates with smaller bias and much smaller variance than than FRHDI with regression fractional weights, specially on the small size of donors. In addition, the variance estimates of fractional regression hot deck imputation (FRHDI) estimators is smaller bias and much smaller variance than those of multiple imputation estimators with the same number of imputations.

In a limited simulation, future work will apply methods to data from longitudinal social science studies, examine more involved simulation contexts, and address situations with multivariate missing data.
Table C.1 Monte Carlo mean and variance of the point estimator under simulation 1

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Method</th>
<th>Mean</th>
<th>Variance</th>
<th>RB(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Complete Sample</td>
<td>2.00</td>
<td>0.0205</td>
<td></td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>FH (M=3)</td>
<td>2.00</td>
<td>0.0738</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>FH+(M=3)</td>
<td>2.00</td>
<td>0.0280</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>MI (M=3)</td>
<td>2.00</td>
<td>0.0280</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>FH (M=5)</td>
<td>2.00</td>
<td>0.0280</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>FH+(M=5)</td>
<td>2.00</td>
<td>0.0270</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>2.00</td>
<td>0.0270</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>FH (M=10)</td>
<td>2.00</td>
<td>0.0264</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>FH+(M=10)</td>
<td>2.00</td>
<td>0.0264</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>2.00</td>
<td>0.0266</td>
<td>0.08</td>
</tr>
<tr>
<td>Domain Mean</td>
<td>Complete Sample</td>
<td>2.00</td>
<td>0.0849</td>
<td></td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>FH (M=3)</td>
<td>2.00</td>
<td>0.1590</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>FH+(M=3)</td>
<td>2.00</td>
<td>0.0843</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>MI (M=3)</td>
<td>2.00</td>
<td>0.0846</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>FH (M=5)</td>
<td>2.00</td>
<td>0.0838</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>FH+(M=5)</td>
<td>2.00</td>
<td>0.0816</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>2.00</td>
<td>0.0818</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>FH (M=10)</td>
<td>2.00</td>
<td>0.0805</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>FH+(M=10)</td>
<td>2.00</td>
<td>0.0803</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>2.00</td>
<td>0.0803</td>
<td>0.08</td>
</tr>
<tr>
<td>$Y &lt; 1$</td>
<td>Complete Sample</td>
<td>0.24</td>
<td>0.0018</td>
<td></td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>FH (M=3)</td>
<td>0.24</td>
<td>0.0052</td>
<td>-0.24</td>
</tr>
<tr>
<td></td>
<td>FH+(M=3)</td>
<td>0.24</td>
<td>0.0028</td>
<td>-0.18</td>
</tr>
<tr>
<td></td>
<td>MI (M=3)</td>
<td>0.24</td>
<td>0.0022</td>
<td>-0.15</td>
</tr>
<tr>
<td></td>
<td>FH (M=5)</td>
<td>0.24</td>
<td>0.0028</td>
<td>-0.23</td>
</tr>
<tr>
<td></td>
<td>FH+(M=5)</td>
<td>0.24</td>
<td>0.0027</td>
<td>-0.23</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>0.24</td>
<td>0.0021</td>
<td>-0.09</td>
</tr>
<tr>
<td></td>
<td>FH (M=10)</td>
<td>0.24</td>
<td>0.0026</td>
<td>-0.15</td>
</tr>
<tr>
<td></td>
<td>FH+(M=10)</td>
<td>0.24</td>
<td>0.0026</td>
<td>-0.26</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>0.24</td>
<td>0.0020</td>
<td>-0.11</td>
</tr>
<tr>
<td>Slope</td>
<td>Complete Sample</td>
<td>1.00</td>
<td>0.0104</td>
<td></td>
</tr>
<tr>
<td>$\theta_4$</td>
<td>FH (M=3)</td>
<td>1.00</td>
<td>0.0882</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>FH+(M=3)</td>
<td>1.00</td>
<td>0.0173</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>MI (M=3)</td>
<td>1.00</td>
<td>0.0167</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>FH (M=4)</td>
<td>1.00</td>
<td>0.0175</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>FH+(M=5)</td>
<td>1.00</td>
<td>0.0164</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>1.00</td>
<td>0.0160</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>FH (M=10)</td>
<td>1.00</td>
<td>0.0161</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>FH+(M=10)</td>
<td>1.00</td>
<td>0.0158</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>1.00</td>
<td>0.0156</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Table C.2  Monte Carlo relative biases, t-statistics and relative variances of the variance estimator under simulation 1

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Method</th>
<th>RB(%)</th>
<th>t-statistic</th>
<th>RV(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>FH− (M=3)</td>
<td>-13.85</td>
<td>-1.41</td>
<td>5421</td>
</tr>
<tr>
<td>θ₁</td>
<td>FH+ (M=3)</td>
<td>-1.46</td>
<td>-0.71</td>
<td>4.62</td>
</tr>
<tr>
<td></td>
<td>MI (M=3)</td>
<td>-1.94</td>
<td>-0.93</td>
<td>10.01</td>
</tr>
<tr>
<td></td>
<td>FH− (M=5)</td>
<td>-1.95</td>
<td>-0.94</td>
<td>5.11</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=5)</td>
<td>-1.65</td>
<td>-0.79</td>
<td>3.42</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>-0.65</td>
<td>-0.32</td>
<td>6.66</td>
</tr>
<tr>
<td></td>
<td>FH− (M=10)</td>
<td>-1.92</td>
<td>-0.92</td>
<td>2.85</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=10)</td>
<td>-2.16</td>
<td>-1.04</td>
<td>2.84</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>-1.99</td>
<td>-0.96</td>
<td>4.16</td>
</tr>
<tr>
<td>Domain Mean</td>
<td>FH− (M=3)</td>
<td>14.96</td>
<td>4.28</td>
<td>5529</td>
</tr>
<tr>
<td>θ₂</td>
<td>FH+ (M=3)</td>
<td>13.25</td>
<td>6.28</td>
<td>18.09</td>
</tr>
<tr>
<td></td>
<td>MI (M=3)</td>
<td>28.32</td>
<td>13.17</td>
<td>27.37</td>
</tr>
<tr>
<td></td>
<td>FH− (M=5)</td>
<td>5.98</td>
<td>2.82</td>
<td>16.85</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=5)</td>
<td>5.75</td>
<td>2.75</td>
<td>14.74</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>29.42</td>
<td>13.72</td>
<td>24.76</td>
</tr>
<tr>
<td></td>
<td>FH− (M=10)</td>
<td>2.56</td>
<td>1.22</td>
<td>14.18</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=10)</td>
<td>2.52</td>
<td>1.21</td>
<td>14.09</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>30.06</td>
<td>14.19</td>
<td>21.42</td>
</tr>
<tr>
<td>Y &lt; 1</td>
<td>FH− (M=3)</td>
<td>6.66</td>
<td>1.36</td>
<td>1997</td>
</tr>
<tr>
<td>θ₃</td>
<td>FH+ (M=3)</td>
<td>-1.76</td>
<td>-0.85</td>
<td>6.38</td>
</tr>
<tr>
<td></td>
<td>MI (M=3)</td>
<td>17.84</td>
<td>8.57</td>
<td>14.61</td>
</tr>
<tr>
<td></td>
<td>FH− (M=5)</td>
<td>-2.26</td>
<td>-1.10</td>
<td>6.07</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=5)</td>
<td>-2.53</td>
<td>-1.24</td>
<td>4.10</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>19.10</td>
<td>9.26</td>
<td>8.57</td>
</tr>
<tr>
<td></td>
<td>FH− (M=10)</td>
<td>-1.91</td>
<td>-0.94</td>
<td>3.10</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=10)</td>
<td>-2.44</td>
<td>-1.19</td>
<td>3.08</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>18.75</td>
<td>9.23</td>
<td>4.91</td>
</tr>
<tr>
<td>Slope</td>
<td>FH− (M=3)</td>
<td>-2.18</td>
<td>-1.29</td>
<td>17321</td>
</tr>
<tr>
<td>θ₄</td>
<td>FH+ (M=3)</td>
<td>0.59</td>
<td>0.28</td>
<td>21.55</td>
</tr>
<tr>
<td></td>
<td>MI (M=3)</td>
<td>2.47</td>
<td>1.14</td>
<td>30.36</td>
</tr>
<tr>
<td></td>
<td>FH− (M=5)</td>
<td>2.92</td>
<td>1.40</td>
<td>31.24</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=5)</td>
<td>0.64</td>
<td>0.31</td>
<td>17.11</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>1.68</td>
<td>0.82</td>
<td>17.25</td>
</tr>
<tr>
<td></td>
<td>FH− (M=10)</td>
<td>3.99</td>
<td>1.94</td>
<td>17.00</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=10)</td>
<td>0.41</td>
<td>0.20</td>
<td>15.36</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>2.10</td>
<td>1.01</td>
<td>13.01</td>
</tr>
</tbody>
</table>
Table C.3 Monte Carlo mean and variance of the point estimator under simulation 2

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Method</th>
<th>Mean</th>
<th>Variance</th>
<th>RB(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Complete Sample</td>
<td>2.00</td>
<td>0.0197</td>
<td></td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>FH−(M=3)</td>
<td>2.00</td>
<td>0.0510</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>FH+(M=3)</td>
<td>2.00</td>
<td>0.0271</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>MI(M=3)</td>
<td>2.04</td>
<td>0.0340</td>
<td>2.00</td>
</tr>
<tr>
<td></td>
<td>FH−(M=5)</td>
<td>2.00</td>
<td>0.0268</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>FH+(M=5)</td>
<td>2.00</td>
<td>0.0260</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>2.04</td>
<td>0.0326</td>
<td>2.01</td>
</tr>
<tr>
<td></td>
<td>FH−(M=10)</td>
<td>2.00</td>
<td>0.0255</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>FH+(M=10)</td>
<td>2.00</td>
<td>0.0256</td>
<td>0.83</td>
</tr>
<tr>
<td>Domain Mean</td>
<td>Complete Sample</td>
<td>2.00</td>
<td>0.0806</td>
<td></td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>FH−(M=3)</td>
<td>2.00</td>
<td>0.1307</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>FH+(M=3)</td>
<td>2.00</td>
<td>0.0839</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>MI (M=3)</td>
<td>2.04</td>
<td>0.0880</td>
<td>2.03</td>
</tr>
<tr>
<td></td>
<td>FH−(M=5)</td>
<td>2.00</td>
<td>0.0824</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>FH+(M=5)</td>
<td>2.00</td>
<td>0.0810</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>2.04</td>
<td>0.0841</td>
<td>2.09</td>
</tr>
<tr>
<td></td>
<td>FH−(M=10)</td>
<td>2.00</td>
<td>0.0786</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>FH+(M=10)</td>
<td>2.00</td>
<td>0.0788</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>2.04</td>
<td>0.0793</td>
<td>2.02</td>
</tr>
<tr>
<td>$Y &lt; 1$</td>
<td>Complete Sample</td>
<td>0.23</td>
<td>0.0018</td>
<td></td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>FH−(M=3)</td>
<td>0.23</td>
<td>0.0059</td>
<td>-0.82</td>
</tr>
<tr>
<td></td>
<td>FH+(M=3)</td>
<td>0.23</td>
<td>0.0029</td>
<td>-0.25</td>
</tr>
<tr>
<td></td>
<td>MI (M=3)</td>
<td>0.23</td>
<td>0.0022</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td>FH−(M=5)</td>
<td>0.23</td>
<td>0.0029</td>
<td>-0.19</td>
</tr>
<tr>
<td></td>
<td>FH+(M=5)</td>
<td>0.23</td>
<td>0.0028</td>
<td>-0.032</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>0.23</td>
<td>0.0021</td>
<td>1.18</td>
</tr>
<tr>
<td></td>
<td>FH−(M=10)</td>
<td>0.23</td>
<td>0.0026</td>
<td>-0.06</td>
</tr>
<tr>
<td></td>
<td>FH+(M=10)</td>
<td>0.23</td>
<td>0.0026</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>0.23</td>
<td>0.0020</td>
<td>1.17</td>
</tr>
<tr>
<td>Slope</td>
<td>Complete Sample</td>
<td>-0.00</td>
<td>0.0598</td>
<td></td>
</tr>
<tr>
<td>$\theta_4$</td>
<td>FH−(M=3)</td>
<td>-0.01</td>
<td>0.0939</td>
<td>77.93</td>
</tr>
<tr>
<td></td>
<td>FH+(M=3)</td>
<td>-0.00</td>
<td>0.0173</td>
<td>11.02</td>
</tr>
<tr>
<td></td>
<td>MI (M=3)</td>
<td>-0.01</td>
<td>0.0852</td>
<td>66.47</td>
</tr>
<tr>
<td></td>
<td>FH−(M=5)</td>
<td>-0.00</td>
<td>0.0683</td>
<td>19.13</td>
</tr>
<tr>
<td></td>
<td>FH+(M=5)</td>
<td>-0.00</td>
<td>0.0672</td>
<td>10.22</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>-0.00</td>
<td>0.0844</td>
<td>8.46</td>
</tr>
<tr>
<td></td>
<td>FH−(M=10)</td>
<td>-0.00</td>
<td>0.0671</td>
<td>18.48</td>
</tr>
<tr>
<td></td>
<td>FH+(M=10)</td>
<td>-0.00</td>
<td>0.0677</td>
<td>12.65</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>-0.00</td>
<td>0.0824</td>
<td>20.16</td>
</tr>
</tbody>
</table>
Table C.4 Monte Carlo relative biases, t-statistics and relative variances of the variance estimator under simulation 2

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Method</th>
<th>RB(%)</th>
<th>t-statistic</th>
<th>RV(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>FH− (M=3)</td>
<td>3.75</td>
<td>0.85</td>
<td>2576</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=3)</td>
<td>2.19</td>
<td>1.07</td>
<td>6.58</td>
</tr>
<tr>
<td></td>
<td>MI (M=3)</td>
<td>11.73</td>
<td>5.36</td>
<td>39.41</td>
</tr>
<tr>
<td></td>
<td>FH− (M=5)</td>
<td>2.53</td>
<td>1.22</td>
<td>6.72</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=5)</td>
<td>3.21</td>
<td>1.56</td>
<td>5.73</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>11.29</td>
<td>5.43</td>
<td>21.99</td>
</tr>
<tr>
<td></td>
<td>FH− (M=10)</td>
<td>3.37</td>
<td>1.66</td>
<td>5.33</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=10)</td>
<td>3.81</td>
<td>1.87</td>
<td>5.66</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>14.38</td>
<td>6.91</td>
<td>16.40</td>
</tr>
<tr>
<td>Domain Mean</td>
<td>FH− (M=3)</td>
<td>19.06</td>
<td>5.23</td>
<td>1384</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=3)</td>
<td>13.07</td>
<td>6.04</td>
<td>31.04</td>
</tr>
<tr>
<td></td>
<td>MI (M=3)</td>
<td>61.83</td>
<td>25.96</td>
<td>92.57</td>
</tr>
<tr>
<td></td>
<td>FH− (M=5)</td>
<td>8.14</td>
<td>3.81</td>
<td>31.04</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=5)</td>
<td>7.13</td>
<td>3.33</td>
<td>30.19</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>62.78</td>
<td>27.90</td>
<td>71.00</td>
</tr>
<tr>
<td></td>
<td>FH− (M=10)</td>
<td>5.81</td>
<td>2.73</td>
<td>31.21</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=10)</td>
<td>5.99</td>
<td>2.81</td>
<td>31.72</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>69.21</td>
<td>31.36</td>
<td>66.61</td>
</tr>
<tr>
<td>Y &lt; 1</td>
<td>FH− (M=3)</td>
<td>8.27</td>
<td>1.37</td>
<td>2420</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=3)</td>
<td>2.64</td>
<td>1.26</td>
<td>7.57</td>
</tr>
<tr>
<td></td>
<td>MI (M=3)</td>
<td>34.95</td>
<td>15.99</td>
<td>34.01</td>
</tr>
<tr>
<td></td>
<td>FH− (M=5)</td>
<td>2.20</td>
<td>1.04</td>
<td>8.97</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=5)</td>
<td>2.40</td>
<td>1.14</td>
<td>4.86</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>32.86</td>
<td>15.47</td>
<td>16.05</td>
</tr>
<tr>
<td></td>
<td>FH− (M=10)</td>
<td>3.56</td>
<td>1.72</td>
<td>3.73</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=10)</td>
<td>2.55</td>
<td>1.23</td>
<td>3.53</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>38.23</td>
<td>18.44</td>
<td>9.74</td>
</tr>
<tr>
<td>Slope</td>
<td>FH− (M=3)</td>
<td>3.68</td>
<td>1.18</td>
<td>2117</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=3)</td>
<td>0.69</td>
<td>0.34</td>
<td>29.41</td>
</tr>
<tr>
<td></td>
<td>FH− (M=5)</td>
<td>1.12</td>
<td>0.55</td>
<td>29.82</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=5)</td>
<td>1.18</td>
<td>0.58</td>
<td>30.38</td>
</tr>
<tr>
<td></td>
<td>MI (M=5)</td>
<td>-19.91</td>
<td>-10.15</td>
<td>18.07</td>
</tr>
<tr>
<td></td>
<td>FH− (M=10)</td>
<td>2.71</td>
<td>1.35</td>
<td>30.90</td>
</tr>
<tr>
<td></td>
<td>FH+ (M=10)</td>
<td>2.26</td>
<td>1.12</td>
<td>32.50</td>
</tr>
<tr>
<td></td>
<td>MI (M=10)</td>
<td>-18.57</td>
<td>-9.42</td>
<td>18.38</td>
</tr>
</tbody>
</table>
BIBLIOGRAPHY


