Weight Adjustments for Fractional Regression Hot Deck Imputation

Minhui Paik and Michael D. Larsen[1]
Iowa State University, Department of Statistics, Ames, Iowa 50011-1210[1], minhui@iastate.edu

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 (FHDI). 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.

KEY WORDS: Calibration; Missing data; Multiple imputation; Quadratic programming; Regression weighting.

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

$$\hat{\theta}_1 = \sum_{i \in A} w_i y_i \tag{1}$$

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 θ_2 is

$$\hat{\theta}_2 = \sum_{i \in A} w_i y_i x_i \tag{2}$$

Hot deck imputation assigns values of y 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

$$y_i^* = \hat{y}_i + \hat{e}_j^* \tag{3}$$

where $\hat{y}_i$ is the predicted value of y_i and $\hat{e}_j^*$ is an imputed residual 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 feature of the method.

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) 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 assigns fractional survey weights to the multiple donors. Fractional imputation was suggested as a method for expressing uncertainty due to imputation and reducing imputation variance. 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 (2006) 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 vary, producing some large weights or even some negative weights. 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 has 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,

$$\begin{aligned} E(y_i|x_i) &= \beta_{0g} + \beta_1 x_i \\ \text{and} \qquad & \\ V(y_i|x_i) &= \sigma^2_g. \end{aligned} \tag{4}$$

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. For $j \in A_{Mg}$, the imputed value for missing y_j is

$$y^*_{Ij} = \sum_{i \in A_{Dgj}} w^*_{ij}(\hat{y}_i + \hat{e}^*_i). \tag{5}$$

In the above formula, $\sum_{i \in A_{Rg}} w^*_{ij} d_{ij} = \sum_{i \in A_{Dgj}} w^*_{ij}$ and A_{Dgj} is the set of indices of imputed values for $j \in A_{Mg}$ The imputed estimator of θ_1 can be constructed as follows:

$$\begin{aligned} \hat{\theta}_{I1} &= \sum_{g=1}^{G} \left[\sum_{i \in A_{Rg}} w_i y_i + \sum_{j \in A_{Mg}} w_j y^*_{Ij} \right] \\ &= \sum_{g=1}^{G} \left[\sum_{i \in A_{Rg}} w_i y_i + \sum_{j \in A_{Mg}} w_j \sum_{i \in A_{Dgj}} w^*_{ij}(\hat{y}_i + \hat{e}^*_i) \right] \\ &= \sum_{g=1}^{G} \left[\sum_{i \in A_{Rg}} w_i y_i + \sum_{j \in A_{Mg}} \sum_{i \in A_{Dgj}} w_j w^*_{ij}(\hat{y}_i + \hat{e}^*_i) \right]. \end{aligned}$$

Kim (2006) suggested adjusting the above formula for weighting under stochastic regression imputation to get an expression for a hot deck imputation. There are two main motivations for Kim's suggestion. First, all missing values can be imputed by observed values. Like hot deck imputation Kim's method just changes the fractional weight to get regression weights 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).

The weighted mean of the imputed values can be written as

$$\sum_{i \in A_{Dgj}} w^*_{ij}(\hat{y}_j + \hat{e}^*_i) = \sum_{i \in A_{Rg}} w^*_{ij} y_i \tag{6}$$

if and only if for each $j \in A_{Mg}$

$$\sum_{i \in A_{Dgj}} w^*_{ij}(1, x_i) = (1, x_j). \tag{7}$$

Therefore, the regression fractionally imputed estimator can be expressed in the form of the fractional hot deck imputed estimator as follows:

$$\hat{\theta}_{I1} = \sum_{g=1}^{G} \left(\sum_{i \in A_{Rg}} w_i y_i + \sum_{j \in A_{Mg}} \sum_{i \in A_{Dgj}} w^*_{ij} y_i \right)$$

where the w_{ij}^* satisfy condition (7).

Similar algebra can be used to write the imputed estimator of θ_2 as a fractional hot deck (FRHDI) estimator. Since $\hat{\theta}_{2I}$ can be expressed as a FHDI estimator, variance estimation through replication can be applied.

3. Construction of Regression Fractional Weights

In order to use fractional regression hot deck imputation, one must construct weights $w_{ij}^*, i \in A_{Rg}$ for each $j \in A_{Mg}$ such that $0 < w_{ij}^* < 1$, $\sum_{i \in A_{Dgj}} w_{ij}^* = w_j$ and formula (7) 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 (2006) suggested the regression weighting method to construct the fractional weights satisfying the constraint (7). 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 α_{ij} and a final fractional weight w_{ij}^* subject to the constraint (7). Let α_{ij} be any initial fractional weights satisfying $\sum_{i \in A_{Dgj}} \alpha_{ij} = 1$. 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 α_{ij} and w_{ij}^* be $Q(\alpha_{ij}, w_{ij}^*) = \sum_{i \in A_{Dgj}} \alpha_{ij}^{-1}(\alpha_{ij} - w_{ij}^*)^2$. Then the problem is to minimize

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

subject to the constraints

$$\begin{array}{ll} & \sum_{i \in A_{Dgj}} w_{ij}^*(1, x_i) = (1, x_j) \\ \text{and} & \\ & 0 < w_{ij}^* < 1, j \in A_{Mg}. \end{array} \qquad (8)$$

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

$$w_{ij}^* = \alpha_{ij} + (x_j - \bar{x}_{Ij}) S_{xx,j}^{-1} \alpha_{ij} (x_i - \bar{x}_{Ij}) \qquad (9)$$

where

$$\begin{array}{lrcl} & S_{xx,j} & = & \sum_{i \in A_{Dgj}} \alpha_{ij} (x_i - \bar{x}_{Ij})^2 \\ \text{and} & & & \\ & \bar{x}_{Ij} & = & \sum_{i \in A_{Dgj}} \alpha_{ij} x_i. \end{array}$$

Under the calibration property $\sum_{i \in A} w_i x_i = \bar{x}_N$, not the full condition (7), 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 (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, including Husain (1969), Bardsley and Chambers (1984), and Rao and Singh (1997). However, the constraint (7) 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 (7), 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 and requires checking the weight at each step against a user supplied criterion. The fractional weight (9) can be rearranged to be

$$w_{ij}^* = \alpha_{ij}(1 + \phi_i) \qquad (10)$$

where

$$\phi_i = (x_j - \bar{x}_{Ij})' S_{xx,j}^{-1} (x_i - \bar{x}_{Ij}).$$

An alternative computational form equivalent to the weights (10) can be constructed as

$$w_{ij}^* = \alpha_{ij} \left(\frac{1 + z_{ij}}{1 + \bar{z}_j} \right), \qquad (11)$$

where

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

and

$$\bar{z}_j = \sum_{i \in A_R} \alpha_{ij} z_{ij}.$$

The regression fractional weights defined by (11) will be nonnegative if $1 + z_{ij} \geq 0$ for every i. Our computer algorithm creates nonnegative fractional weights by modifying the z_{ij} such that $|z_{ij}| < 1$ so that $w^*_{ij} > 0$ and there is an L such that $max_{i \in A_{Dgj}} (w^*_{ij}) < L$. If the first-step fractional weights fall outside the desired range $|z_{ij}| < 1$, then 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 α_{ij} are used as first-step weights. Adjustment values can be obtained by γ_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

$$z_{ij} = (x_j - \bar{x}_{Ij})' \left(\sum_{i \in A_{Dj}} \alpha_{ij} (x_i - x_j)^2 \right)^{-1} (x_i - x_j)$$

and

$$\bar{z}_j = max\{ \sum_{i \in A_{Dj}} \alpha_{ij} z_{ij}, \frac{1}{M} - 1 \}$$

If $|z_{ij}| < 1$ for every $i \in A_{Dj}$, then the initial weights α_{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

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

and

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

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

STEP 3: Compute the new regression fractional weights:

$$\lambda_i^{(k)} = \prod_{j=1}^{(k)} \gamma_i^{(j)}$$

$$z_{ij}^{(k)} = (x_j - \bar{x}_{Ij})' \left(\sum_{i \in A_{Dj}} \alpha_{ij} \lambda_i^{(k)} (x_i - x_j)^2 \right)^{-1} \times (x_i - x_j)$$

$$\bar{z}_j^{(k)} = max\{ \sum_{i \in A_{Dj}} \alpha_{ij} z_{ij}^{(k)}, \frac{1}{M} - 1 \}$$

STEP 4: If $|z_{ij}^{(k)}| < 1$ for every $i \in A_{Dj}$, then

$$w_{ij}^{*(k)} = \alpha_{ij} \left(\frac{1 + z_{ij}^{(k)}}{1 + \bar{z}_j^{(k)}} \right).$$

If not, set $k = k + 1$ and go to STEP 2.

The final fractional weights $w_{ij}^{*(s)}$ have the following properties. For each $j \in A_{Mg}$,

(i) $0 < w_{ij}^{*(k)} < 1$ for $i \in A_{Dgj}$ and

(ii) $\sum_{i \in A_{Dgj}} w_{ij}^{*(k)} (1, x_i) = (1, x_j)$.

If the imputation cells are such that the restriction cannot be met, the program will produce weights approximating the criterion. In our situation, we can always find nonnegative fractional weights by the proposed computer algorithm 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.

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:

FRHDI0 : FRHDI using the regression fractional weight,

FRHDI1 : FRHDI using the nonnegative fractional weight,

and

MI : Multiple imputation.

For fractional imputation, for each missing value, we created 10-case nearest neighborhoods where the distance is defined on the value of x_i. Each nearest neighborhood consists of 10 respondents with the closest x-value to the value x_j for missing unit j. After that, M donors for each missing unit are selected by simple random sampling without replacement within each nearest neighborhood. Fractional weights w^*_{ij} on FRHDI0 are calculated by the regression method (9) setting the initial fractional weight equal to $\alpha_{ij} = 1/d_{ij}$. Nonnegative fractional weights are obtained by the computer algorithm that was described in Section 3.2. For multiple imputation (Rubin 1987), M repeated imputation values are drawn from the posterior predictive distribution of the missing values under a simple linear regression model given the standard prior distribution, e.g., constant on the regression coefficients and inversely proportional to the regression model variance. In this simulation, we used $M = 5$ 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

θ_1 = mean of Y,

θ_2 = slope of Y on X, and

θ_3 = mean of Y where $z < 0.25$.

For fractional imputation, the variance estimation method proposed by Kim (2006) was applied. Kim (2006) considers the adjusted jackknife replicates constructed by decreasing by an appropriate amount fractional weights of the imputed values associated with a deleted respondent and increasing by an appropriate amount 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 FRHDI0. The variance estimator for multiple imputation was given in Rubin (1978, 1987).

Table 1 shows the mean and variance of the imputed estimator calculated based on the Monte Carlo samples generated by the linear regression model. The three imputation methods are unbiased. Fractional imputation is slightly more efficient than multiple imputation with the same number of donors. In addition, there is no further improvement using FRHDI with nonnegative

Table 1: Monte Carlo mean and variance of the point estimator under simulation 1.

Parameter	Method	Mean	Variance
Mean(θ_1)	FRHDI0(M=5)	2.00	0.0262
	FRHDI1(M=5)	2.00	0.0260
	MI (M=5)	2.00	0.0268
	FRHDI0(M=10)	2.00	0.0250
	FRHDI1(M=10)	2.00	0.0260
	MI (M=10)	2.00	0.0260
Slope(θ_2)	FRHDI0(M=5)	1.00	0.0158
	FRHDI1(M=5)	1.00	0.0155
	MI (M=5)	1.00	0.0162
	FRHDI0(M=10)	1.00	0.0156
	FRHDI1(M=10)	1.00	0.0158
	MI (M=10)	1.00	0.0159
Domain mean(θ_3)	FRHDI0(M=5)	2.00	0.0778
	FRHDI1(M=5)	2.00	0.0765
	MI (M=5)	2.00	0.0830
	FRHDI0(M=10)	2.00	0.0781
	FRHDI1(M=10)	2.00	0.0805
	MI (M=10)	1.99	0.0790

Table 2: Monte Carlo relative biases and t-statistics of the variance estimator under simulation 1.

Parameter	Method	RB(%)	t-statistic
Mean(θ_1)	FRHDI0(M=5)	-1.74	-0.86
	FRHDI1(M=5)	-0.71	-0.35
	MI (M=5)	3.54	1.81
	FRHDI0(M=10)	3.25	1.63
	FRHDI1(M=10)	-0.14	-0.07
	MI (M=10)	5.13	2.51
Slope(θ_2)	FRHDI0(M=5)	3.87	0.04
	FRHDI1(M=5)	2.20	0.02
	MI (M=5)	5.73	0.07
	FRHDI0(M=10)	1.78	0.02
	FRHDI1(M=10)	1.54	0.02
	MI (M=10)	3.39	0.04
Domain mean(θ_3)	FRHDI0(M=5)	7.60	3.71
	FRHDI1(M=5)	4.62	2.30
	MI (M=5)	28.48	13.82
	FRHDI0(M=10)	5.61	2.62
	FRHDI1(M=10)	2.23	1.07
	MI (M=10)	32.85	15.40

fractional weights based on the efficiency of the point estimator.

In Table 2, relative biases and t-statistics 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 fractional imputation variance estimation procedures 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 2 further illustrates that multiple imputation produces a seriously biased estimator of the variance of the estimator of θ_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 or 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.

One additional result can be mentioned. The variance estimators for FRHDI1 are more stable than those for FRHDI0 and MI. FRHDI0 has the uniformly smallest variance of the variance estimators on all three parameters.

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, 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 estimand parameters we consider in simulation 2 are the same as in simulation 1.

Table 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 are approximately unbiased and have more efficiency than MI for

Table 3: Monte Carlo mean and variance of the point estimator under simulation 2.

Parameter	Method	Mean	Variance
Mean(θ_1)	FRHDI0(M=5)	2.01	0.0265
	FRHDI1(M=5)	2.01	0.0264
	MI (M=5)	2.04	0.0340
	FRHDI0(M=10)	2.01	0.0259
	FRHDI1(M=10)	2.01	0.0263
	MI (M=10)	2.04	0.0322
Slope(θ_2)	FRHDI0(M=5)	0.00	0.0655
	FRHDI1(M=5)	0.00	0.0649
	MI (M=5)	0.55	0.0096
	FRHDI0(M=10)	0.00	0.0645
	FRHDI1(M=10)	0.00	0.0681
	MI (M=10)	0.54	0.0085
Domain mean(θ_3)	FRHDI0(M=5)	2.01	0.0841
	FRHDI1(M=5)	2.01	0.0822
	MI (M=5)	2.04	0.0921
	FRHDI0(M=10)	2.02	0.0781
	FRHDI1(M=10)	2.01	0.0779
	MI (M=10)	2.04	0.0840

Table 4: Monte Carlo relative biases and t-statistics of the variance estimator under simulation 2.

Parameter	Method	RB($\%$)	t-statistic
Mean(θ_1)	FRHDI0(M=5)	2.49	1.25
	FRHDI1(M=5)	1.47	0.72
	FRHDI0(M=10)	2.98	1.44
	FRHDI1(M=10)	1.04	0.52
Slope(θ_2)	FRHDI0(M=5)	4.79	1.84
	FRHDI1(M=5)	3.87	1.56
	FRHDI0(M=10)	6.59	2.50
	FRHDI1(M=10)	0.77	0.30
Domain mean(θ_3)	FRHDI0(M=5)	7.13	3.38
	FRHDI1(M=5)	4.63	2.23
	FRHDI0(M=10)	7.67	3.65
	FRHDI1(M=10)	6.38	3.08

all parameters. 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 especially for the slope. This result indicates that, as expected, MI is very sensitive to the model used for imputation. Bias of multiple imputation estimator could be improved to some degree if the imputed values were generated by the local simple linear regression model or by a quadratic regression model.

In Table 4 the biases and t-statistics of the two fractional imputation methods are illustrated under simulation 2. The simulation results of multiple imputation variance estimators are not listed in Table 2 since the point estimators are seriously biased. Based on the biases and t-statistics in Table 4, FRHDI variance estimation procedures are unbiased for θ_1 and θ_2. The variance estimator using nonnegative fractional weights (FRHDI1) is more stable than FRHDI0.

5. Summary and Discussion

We have discussed a procedure for constructing nonnegative adjusted fractional weights for fractional regression hot deck imputation (FRHDI). In some situations, of course, solutions to the algorithm do not exist. Future work will examine options when, for example, donor values with a spread of x-values are not available.

In a limited simulation, the proposed method performs better than naive multiple imputation and FRHDI without the restrictions on the fractional weights. Future work will apply methods to data from longitudinal social science studies, examine more involved simulation contexts, and address situations with multivariate missing data.

ACKNOWLEDGMENTS

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

REFERENCES

Bardsley, P., and Chambers, R.L., (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

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

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

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

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

Huang, E.T., and Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics section*, American Statistical Association, 57-64.

Husain, M. (1969). Construction of Regression weights for estimation in Sample Surveys. Unpublished M.S. thesis, Iowa State University, Ames, Iowa.

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

Kim, J.K. (2006). Fractional Regression Hot Deck Imputation. Draft manuscript.

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

Rao, J.N.K., and Singh, A.C. (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. *Proceedings of the section on survey research methods*, American Statistical Association, 57-64.

Rubin, D.B. (1978). Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse. *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association (Alexandria, VA), 20-28.

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