Fractional Regression Nearest Neighbor Imputation

Minhui Paik and Michael D. Larsen
Department of Statistics, Snedecor Hall, Ames, Iowa 50011, paik@iastate.edu, larsen@iastate.edu

Abstract

Sample surveys typically gather information on a sample of units from a finite population and assign survey weights to the sampled units. Survey frequently have missing values for some variables for some units. Fractional regression imputation creates multiple values for each missing value by adding randomly selected empirical residuals to predicted values. Fractional imputation methods assign fractional survey weights to the imputed values. Fractional nearest neighbor imputation randomly selects multiple donors for each missing value from a set of nearest neighbors. The fractional regression nearest neighbor imputation method developed in this paper imputes more than one value for each missing item using donors that are neighbors selected by a distance calculation involving both regression model predictions and variables used in other nearest neighbor methods. Different distance function specifications, which can involve both observed and predicted values, produce alternative imputation procedures. In this paper, we compare the performance of fractional imputation methods, including fractional regression nearest neighbor imputation, in a simulation study. In addition, we examine empirically the performance of the imputation methods studied in this paper on a subset of data from the Iowa Family Transitions Project under different missing data assumptions.

Keywords: Cell mean model, Hot deck, Missing data, Multiple imputation, Regression imputation.

1. Introduction

Sample surveys typically gather information on a sample of units from a finite population and assign survey sampling weights to the sampled units. Surveys usually have missing values for some variables for some units. Imputation methods fill in the missing values with plausible values to create a completed data set. Various imputation methods have been developed to compensate for item nonresponse. Simple imputation methods are commonly used in practice, but these may not be adequate in many circumstances. More sophisticated methods, such as fractional hot deck imputation and multiple imputation, have been developed and may be preferable for representing uncertainty due to imputation. Methods that are more or less parametric in terms of imputation model are available.

Fractional imputation was developed to reduce the imputation variance which came from the random component of the variance of the estimator arising from imputation. Kim and Fuller (2004) investigated the method of fractional hot deck for the cell mean response model. Fractional imputation using a hot deck method was shown to have certain advantages related to avoiding distributional assumptions that arise in parametric multiple imputation. In their study, the fractional hot deck imputation is more efficient than multiple imputation based on the same number of donors. The authors also suggest a consistent replication variance estimation procedure for their fractional hot deck method. Hot deck imputation of a single missing value using the cell mean model can not, however, 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.

Regression imputation uses covariates to predict missing values. Regression imputation is potentially advantageous if the variable of interest is strongly related to auxiliary variables. Kim (2003) studied fractional imputation using a regression imputation model. Fractional regression imputation creates multiple values for each missing value by adding randomly selected empirical residuals to predicted values. Each imputed value is assigned a fractional survey weight.

Fractional nearest neighbor imputation also was studied by Fuller and Kim (2005a,b). They demonstrated the properties of the estimator under some conditions and developed a jackkife variance technique for fractional nearest neighbor imputation.

In this paper, we extend the fractional nearest neighbor imputation to fractional regression nearest neighbor imputation. The new method uses a suitable distance measure to choose nearest neighbors. It preserves the correlation structure among quantitative variables, thereby combining the advantages of both a nearest neighbor method and a regression model using fractional imputation. Since this new procedure is a specific case of fractional nearest neighbor imputation, a jackknife variance estimation technique developed by Fuller and Kim (2005a,b) can be applied for variance estimation. A simulation is conducted to compare imputation methods.

Methods are applied to a subset of data from the Iowa Family Transitions Project (IFTP), which is a combination of the Iowa Youth and Family Project (IYFP) and the Iowa Single Parent Project (ISSP). The Iowa Family Transitions Project (IFTP) is one of those rare opportunities when researchers have repeatedly entered the lives of individuals to chart the development of their relationships, and link variability in relationship quality to antecedent conditions in families of origin and to important consequences such as physical health and emotional well being. The FTP involves the study of a cohort of over 500 young adults that began in 1989 and has continued for the past 15 years. The original project had its genesis in the rural ”farm crisis” of the late 1980s. One of its central

objectives was to document the effects of family adversity on the physical, emotional and behavioral health of adolescents. Publications based on this data set include Conger et al (1990), Lorenz et al (1991), Wickrama, Conger, and Lorenz (1995), and Conger, Lorenz and Wickrama (2004). Missing data in this data set decrease its power for detecting significant statistical relationships. The relative performances of different imputation methods are compared.

This article is organized as follows. In section 2, the properties of several imputation methods that are related to fractional regression imputation are discussed. In section 3, fractional regression imputation is described and the problem of estimating the variance of the estimator is addressed briefly. In section 4, fractional regression nearest neighbor imputation is defined and studied. A simulation study in section 5 evaluates fractional regression imputation, fractional regression nearest neighbor imputation, and other imputation methods with regards to efficiency of a point estimator. In section 6, the imputation methods described in this paper are implemented for a subset of the IFTP data. A summary and discussion of future work are given in section 7.

2. Imputation methods

2.1 Nearest Neighbor Imputation

Nearest neighbor imputation (NNI) selects the respondent closest to the non-respondent by minimizing a specified 'distance' (Kalton 1983, Lessler and Kalsbeek 1992, Rancourt 1999, Rancourt, Särndal, and Lee 1994, Chen and Shao 2000 and 2001) and its value is substituted for the nonrespondent. Chen and Shao (2000) summarize some of the advantages of the NNI method. First, the missing items are replaced by the observed units so that the imputed values are actually observed values, not constructed values. Second, since the NNI method used the information of the auxiliary variables, the NNI method may be more efficient than other hot deck imputation schemes. Third, it makes no distributional assumptions in comparison to the explicit models such as regression imputation using a normal linear regression model. Chen and Shao (2000) prove that the nearest neighbor approach estimates distributions correctly under some conditions, but produces bias if these conditions are not met. In particular, a skew distribution can lead to bias.

NNI can be criticized, however, because it imputes only a single value for each missing value. Therefore, it cannot represent uncertainty due to imputation without special variance estimation formulas, such as those in Chen and Shao (2000, 2001).

The estimate of the mean of a variable Y is the sample mean of the observed plus imputed values. The estimate of a regression slope for a prediction of a variable Y from a variable X is the standard least squares regression estimate using the (x, y) pairs, where some y-values are observed and the others are imputed.

2.2 Stochastic Regression Imputation

Another classical method for imputing missing data is (stochastic) regression imputation. In this method, a missing value is replaced by a value predicted by regression imputation plus a residual, drawn to reflect uncertainty in the predicted value. This residual can be obtained by two alternative ways: i) by drawing from a normal distribution with mean zero and and estimated standard deviation, or ii) by selecting randomly from the set of empirical residuals. Method ii) is preferred when some assumptions of the regression model are not reasonable. In ii), if few values are missing, then sampling of residuals can be done without replacement. If, however, a relatively large fraction of values are missing, then sampling can be done with replacement from the observed residuals.

Stochastic regression imputation maintains the distribution of the variables in the sense of maintaining the observed relationship between a variable Y and its predictor variables and allows for the estimation of distributional quantities (Kalton and Kasprzyk 1982, Kalton 1983, Nordholt 1998). However, such a parametric approach is potentially more sensitive to model violations than methods based on implicit models. If the regression model is not a good fit, then the predictive power of the model might be poor (Little and Rubin 2002). In addition, the imputed value is the predicted value plus a residual, which is not an actually occurring values. In case of certain types of variables such as earnings and income variables, that fact could be a problem. Further, single imputation cannot represent uncertainty due to imputation unless special formulas, such as those of Rao and Shao (1992; see also Rao 1996) are used.

Estimates of the mean of Y and of a regression coefficient are computed as they were with NNI.

2.3 Fractional Imputation

The method of fractional imputation (FI) was originally suggested as a method for improving the efficiency of the imputed point estimator by eliminating variance (conditional on an observed sample) due to imputation. FI, suggested by Kalton and Kish (1984) and studied by Kim and Fuller (2004), selects multiple donors for each missing observation and assigns a weight equal to a fraction of the original survey weight for each donor. In a cell mean model (equal mean within a cell), donors are selected from within the cell.

Fully efficient fractional imputation (FEFI) uses all observed cases within a cell as donors for the missing cases. Kim and Fuller (2004) found for the cell mean model that FI and their variance estimator are superior to multiple imputation (MI; Rubin 1978; see also Rubin 1987, 1996) estimators using a parametric model based on the same number of multiple donors. The improvement can be explained by the fact that MI adds additional variability due to the drawing of parameters from their posterior distribution to the variance of the imputation-based estimator.

Durrant (2005) mentioned that one potential advantage of FI is that multiple data sets may not need to be stored which could make the data handling under FI under certain circumstances easier than under MI where M completed data files need to be stored and analyzed. Under fractional hot deck imputation, it is enough to store only the replication weights that indicate how often a donor has been used for imputation to carry out further analysis (Kim and Fuller 2004).

The estimator of the mean of Y from the fractionally imputed sample is $\sum_{i\in s_R}\sum_{j\in s} y_i w_{ij}^*$, where $w_{ij}^* = M^{-1}d_{ij}$ is the fraction of the weight allocated to donor i for recipient j, M is the number of imputations for a missing y-value, s_R is the set of indices of the respondents, and s is the set of indices for the full sample. In the above formula, if unit j is observed ($j \in s_R$), $d_{jj} = 1$ and $d_{ij} = 0$ for $i \neq j$. The estimator of the slope coefficient in a simple linear regression is $\sum_{i\in s_R}\sum_{j\in s}(x_j - \bar{x})(y_i - \bar{y}_I)w_{ij}^* / \sum_{j\in s}(x_j - \bar{x})^2$, where $\bar{y}_I$ is the weighted mean of the imputed sample. That is, the regression slope estimator is the weighted least squares estimator.

3. Fractional Regression Imputation

The method of fractional regression imputation (Kim 2003) is a composite method defined by combining ideas from fractional imputation and regression imputation. The reason for utilizing a regression imputation method is to preserve the correlation structure between an outcome variable and predictor variables. Hot deck imputation, at least under the cell means models, ignores quantitative auxiliary predictor variables.

The fractional regression estimator is as follows. First, compute the regression of y on x by classical least squares using the pairs (x, y) with an observed y-value and estimate the missing y-values by the estimated regression function. Second, compute the residuals for the observed y-values. Randomly draw M residuals without replacement sampling for each missing value. Third, insert the predicted value plus M residuals for each missing value and assign a weight M^{-1} to each of the imputed values. Fourth, the estimator of the mean of y and the regression coefficient for the regression of y on x are computed as described previously for fractional imputation. That is, for the slope, weighted least squares is used.

In order to decrease the impact of the standard normal linear regression assumptions on imputations, one can choose residuals randomly within imputation cells. One method of defining imputation cells that should be responsive to some deviations from standard assumptions, such as heterogeneous error variances and slight curvature of the $X - Y$ relationship, is to define imputation cells by intervals of the X variable. In practice, this can be accomplished by dividing the X-range into G intervals. In the simulation, the intervals are defined by equally spaced quantiles because the distribution used for X is quite skew. In the case of ten intervals in the simulation, if one cell had no observed values, then the pool of donors was expanded to include observed cases from neighboring cells. In other applications, one could use equally spaced intervals or a customised interval construction.

Kim (2003) studied the properties of point estimators computed with imputations generated with fractional regression imputation. The estimators are unbiased for the marginal mean of Y under models and response mechanisms considered in Kim (2003). Fractional regression imputation is a variation on stochastic regression imputation that uses fractional imputation to reduce the variance due to imputation. It does this by selecting multiple residuals randomly from the set of empirical residuals.

Kim (2003) suggested a replication variance estimation method. This suggested variance method is not desirable to use in practice because the variance is calculated by fully efficient estimator in which all respondent residuals are used as donated residuals for each missing value. Since the imputed value resulting from fractional regression imputation is not actually an observed value, the replication variance method suggested by Kim and Fuller (2003) does not work for variance estimation. The method of Kim and Fuller (2003) uses the idea of replication weights to produce a variance estimate; these require the use of only observed values.

It is possible to modify a jackknife variance estimation technique for variance estimation under fractional regression imputation. This approach does not appear to have been explored yet. That subject will be a topic of future research. In simulations reported in this paper, we simply use the replication variance method suggested by Kim and Fuller (2003) through considering the respondent which gives a residual to the recipient as a donor. That is, the jackknife is applied to units rather than values.

4. Fractional Regression Nearest Neighbor Imputation

Fractional regression nearest neighbor imputation is developed in this paper to achieve the advantages of both a nearest neighbor method and a regression model using fractional imputation. In order to combine both methods using fractional imputation, a natural starting place is to find the nearest neighbour donors which preserve the correlation structure. First, compute the regression of y on x by classical least squares and estimate the missing y-values by the estimated regression function. Draw a residual randomly without replacement for each missing value. Second, let the pseudo imputed value be the predicted value plus the residual for each missing value. Third, compute the distance from all respondents to the pseudo imputed value. The missing values is replaced by the value that has the minimum distance to the imputed point. The process repeated M times for each missing value to created M imputations. Estimation based on the fractional imputed data set is performed analogously to the method with fractional regression imputation. As such, the resulting procedure and estimator can be viewed as a specific case of a replicated nearest neighbor estimator.

Fractional regression nearest neighbor imputation is similar in spirit to predictive mean matching (Little 1988) and a regression-based nearest neighbor hot deck procedure (Laaksonen 2000). The innovation with the method of this paper is to combine the ideas with fractional imputation to reduce variance due to imputation.

Fuller and Kim (2005) studied the model and estimator properties of replicated nearest neighbor imputation and outlined a replication variance estimator closely related to that of Kim and Fuller (2004). The variance estimator changes the fractional replicate weights of the naive variance estimator to produce a consistent estimator of the variance. Their variance estimation procedure should be usable with this new version of nearest neighbor fractional imputation.

In detail, to get M donors, the procedure for fractional regression nearest neighbor imputation as follows.

1. STEP1 Compute the regression line and observed residuals.
   (a) Compute the regression of Y on X using least squares estimation based on the observed data pairs (x, y): $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.
   (b) Compute predicted values for all points in the sample: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
   (c) Compute residuals for all points in the sample with observed y-values: $\hat{e}_i = y_i - \hat{y}_i$.
2. STEP2 For each observation with a missing value of y, compute a pseudo imputed value $y_i^* = \hat{y}_i + \hat{e}^*$ as follows:
   (a) Randomly select a residual $\hat{e}^*$ from the set of observed residuals: $\{\hat{e}_j = y_j - \hat{y}_j; j \in s_R\}$.
   (b) Let $y_i^* = \hat{y}_i + \hat{e}^*$.
3. STEP3 Find the donor y_i^{**} for the missing case i.
   (a) Compute the distance between y_i^* and all points with observed y-values: $\mid y_j - y_i^* \mid$.
   (b) Select the case j that produces the minimum distance: $\min_{j \in A_R} \mid y_j - y_i^* \mid$.
   (c) Let the imputed value for case i be $y_i^{**} = y_j$, where j is determined by the previous step.
4. STEP4 Repeat steps 2-3 M times and M different donors are chosen for each missing value.

5. Monte Carlo Study

A simulation study was conducted to evaluate fractional regression nearest imputation and other imputation methods with regards to efficiency of a point estimator. Two variables were generated. Independent variable X_i was generated from a chi-squared distribution with one degrees of freedom. Response variable Y_i was generated as $Y_i = 2 + 0.5X + e_i$, where $e_i \sim N(0, 1)$ independent from X_i. The correlation between X and Y is 0.58. The response indicator variable R_i is generated from a Bernoulli distribution with the response rate $p = 0.65$. That is, the data are missing completely at random in this simulation. Future simulations will consider other probability mechanisms for missing data. We generated $B = 10000$ replicate samples of size $n = 100$.

Imputation cells for fractional regression imputation are formed using the values on the X variable. Cells were formed based on equally-spaced quantiles of X. The impact of the number of cells was studied. That is, the data in each sample were divided into $G = 1, 3, 5,$ or 10 cells for separate analyses.

The following methods are compared in the simulation.

1. CC Complete cases analysis.
2. NN Nearest neighbor matching on X.
3. SR Single imputation stochastic regression. Method 1 (SR1) draws residuals from an estimated distribution. Method 2 (SR2) randomly selects empirical residuals.
4. MI Multiple imputation with $(M = 5)$ imputations under the normal linear regression model with a prior distribution proportional to the inverse of the regression error variance; i.e., the standard noninformative prior distribution (Gelman et al 2004; section 14.2).
5. FRI Fractional regression imputation with $(M = 5)$ imputations. The imputation cells are formed by using similar X-values. The number of cells is denoted as FRI1, FRI3, FRI5, and FRI10.
6. FNNI Fractional nearest neighbor imputation with $(M = 5)$ imputations. The M respondents closest to the value of a missing X are selected as donors.
7. MRNNI Multiple regression nearest neighbor imputation with $(M = 5)$ imputations. The imputation cells for creating pseudo-imputations are formed by using similar X-values. The number of cells is denoted as MRNNI1, MRNNI3, MRNNI5, and MRNNI10. Instead of weighting donors, values are multiply imputed and MI combination formulas are used.
8. FRNNI Fractional regression nearest neighbor imputation with $(M = 5)$ imputations. The imputation cells for creating pseudo-imputations are formed by using similar X-values. The number of cells is denoted as FRNNI1, FRNNI3, FRNNI5, and FRNNI10.

Results are reported for point estimators of the mean of Y and for the slope of the regression of Y on X. Future work will complete evaluation of these methods and also study estimation of variances.

Table 1: Monte Carlo means and standard deviations of point estimators of the mean of the outcome variable and of the linear regression slope based on 10000 replications. Variable X is chisquared(1) and $n = 100$.

	Mean of Y		Slope	
Method	Mean	SD	Mean	SD
CC	2.501	0.151	0.501	0.098
NN	2.495	0.156	0.483	0.111
SR1	2.501	0.154	0.501	0.108
SR2	2.501	0.153	0.501	0.108
MI	2.501	0.145	0.501	0.102
FRI1	2.501	0.144	0.501	0.100
FRI3	2.501	0.144	0.501	0.101
FRI5	2.501	0.145	0.501	0.102
FRI10	2.501	0.147	0.501	0.102
FNNI	2.484	0.144	0.460	0.104
MRNNI1	2.500	0.144	0.487	0.099
MRNNI3	2.496	0.145	0.487	0.099
MRNNI5	2.497	0.145	0.487	0.099
MRNNI10	2.497	0.146	0.489	0.100
FRNNI1	2.494	0.144	0.477	0.097

Table 2: Monte Carlo means and standard deviations of point estimators of the mean of the outcome variable and of the linear regression slope based on 10000 replications. Variable X is normal and $n = 100$.

	Mean of Y		Slope	
Method	Mean	SD	Mean	SD
CC	2.501	0.153	0.501	0.091
NN	2.501	0.157	0.495	0.101
SR1	2.501	0.156	0.501	0.100
SR2	2.501	0.155	0.502	0.099
MI	2.501	0.146	0.502	0.094
FRI1	2.500	0.145	0.501	0.093
FRI3	2.501	0.146	0.501	0.093
FRI5	2.500	0.146	0.501	0.093
FRI10	2.501	0.148	0.502	0.094
FNNI	2.501	0.146	0.484	0.090
MRNNI1	2.501	0.145	0.496	0.091
MRNNI3	2.500	0.145	0.497	0.092
MRNNI5	2.501	0.146	0.498	0.092
MRNNI10	2.500	0.148	0.499	0.093
FRNNI1	2.501	0.145	0.495	0.090

The mean of variance of the point estimators of the response variable grand mean and the regression slope coefficient were calculated based on $B = 10000$ simulations. Table 1 shows the means and standard deviations of the point estimators under various methods. Methods CC, SR, MI, and FRI were unbiased for the slope and the mean of Y, but those that rely on picking nearest neighbors (NN, FNNI, MRNNI, FRNNI) were not. These methods tend to underestimate the slope and mean of Y. The principle cause of this bias is that the distribution used to generate the x-values: chi-square with one degree of freedom. When a y-value is missing and a nearest neighbor in the X-dimension is chosen, slightly more than half the time (53% in simulations) the nearest neighbor is below the actual x-value. Since there is a positive correlation of X with Y, the donated y-values then are slightly less on average than the real values. This depresses the estimated mean of Y and the estimated regression slope. In FRNNI, which matches on Y, the nearest neighbors in the Y-dimension also slightly more than half the time are below the actual y-value. The effect is not large, but it is noticeable.

The Monte Carlo standard deviation of Table 1 shows that methods based on imputing multiple values and taking the average (MI, FRI, FNNI, MRNNI, FRNNI) produce smaller variation in point estimates than complete case analysis (CC) and methods imputing a single value (NN, SR). The number of classes in FRI (and, we believe, in FRNNI) do not make much difference, but the standard deviation of point estimates increases slightly as the number of classes increases. The apparent bias also seems to decrease slightly as the number of classes increases. Single imputation methods add variability above that of the complete case analysis; this is variability due to imputation.

A second simulation was conducted with values of X generated from a normal distribution with mean 1 and variance 2, which match those of the chisquare distribution with one degree of freedom. The correlation of X with Y is 0.58 as before. In the second simulation, the effect of nearest neighbor matching is reduced, but not totally eliminated. Matching X-values tend to be more in the center of the X-distribution, so donated Y values tend to be closer to the mean. This does not cause bias in the estimate of the mean of Y, but it does depress the apparent slope in the regression relationship. That is, large values of X are matched more often than 1/2 with values that are closer to the mean of X, which tend to have smaller y-values than the real values that are missing. Similarly, small values of X also are matched more often than 1/2 with values that are closer to the mean of X, which tend to have larger y-values than the real values that are missing. Other patterns of results, which are presented in Table 2, are similar as those in the first simulation. The standard deviations of slope point estimates are lower in Table 2 than in Table 1 due to the influence of the skewness of X for Table 1 and symmetry in Table 2.

A third simulation was run to study the effect of sample size. Variable X was again generated from a chisquare(1) distribution, but sample size was increased to $n = 200$. As seen in Table 3, the biases of the mean and slope estimators are reduced, but not eliminated. Other results are the same as before.

Table 3: Monte Carlo means and standard deviations of point estimators of the mean of the outcome variable and of the linear regression slope based on 10000 replications. Variable X is chisquare(1) and $n = 200$.

	Mean of Y		Slope	
Method	Mean	SD	Mean	SD
CC	2.499	0.108	0.500	0.065
NN	2.497	0.111	0.489	0.075
SR1	2.500	0.118	0.500	0.078
SR2	2.500	0.109	0.500	0.072
MI	2.500	0.106	0.500	0.070
FRI1	2.500	0.103	0.500	0.066
FRI3	2.500	0.103	0.500	0.067
FRI5	2.500	0.103	0.500	0.067
FRI10	2.500	0.104	0.500	0.067
FNNI	2.491	0.103	0.474	0.071
MRNNI1	2.500	0.103	0.491	0.066
MRNNI3	2.500	0.102	0.491	0.067
MRNNI5	2.500	0.103	0.491	0.067
MRNNI10	2.500	0.104	0.492	0.067
FRNNI1	2.494	0.144	0.477	0.097

6. Performance of the Methods on the IFTP Data

The dataset used to study imputation methods is based on the Iowa Youth and Family Project (IYFP) and the Iowa Single Parent Project (ISSP). These studies are part of a long-term sociology project, which started in the 1990s, after the farm crisis of the late 1980s in rural Iowa. The aim was to observe and analyze the changing dynamics of families due to the financial hardships suffered, and the result on the relationships between the different members of the family. This data set is chosen, because item missing data and drop-out over time reduce the statistical power of statistical analyses. The ultimate goal is to be able to recommend general missing data methods for use in this data set so that various analyses can be conducted utlizing all available information, including cases with complete and parital response.

One subject that the sociologists and psychologists involved in the project wanted to study is the impact of economic hardship on the self-esteem (SE) of the subjects. Economic hardship is viewed as a latent variable. Therefore, the researchers designed a list of questions that helped them quantify and observe three other variables, namely financial strain (FINSTR), making ends meet (ENDS) and financial cutbacks (CUT), which in turn estimated the economic hardship suffered by the subject family. Substantive analyses utilize structural equation models based on four waves of data (1991, 1992, 1994, and 2001).

In the present work, we study the impact of missing data methods on estimates of the relationship between financial strain (QFFINSTR; the X variable) and self esteem (QFSE; the Y variable) and the mean of QFSE in 1991. The particular interest of this study is a comparison of fractional regression nearest neighbor imputation and other imputation methods with regards to efficiency of a point estimator under different missing data mechanisms. Missing data mechanisms are discussed below.

Table 4: Monte Carlo means and standard deviations of point estimators of the mean of the outcome variable and of the linear regression slope based on 10000 replications. Variable X is QFFINSTR, Y is QFSE, $n = 100$, and the missing data mechanism is MCAR.

	Mean of Y		Slope	
Method	Mean	SD	Mean	SD
Full data	4.000		-0.2410	
CC	4.000	0.0549	-0.2419	0.0816
NN	4.003	0.0604	-0.2099	0.0833
SR1	4.000	0.0600	-0.2420	0.0889
SR2	3.999	0.0599	-0.2422	0.0886
MI	4.000	0.0621	-0.2418	0.0845
FRI1	3.999	0.0548	-0.2417	0.0832
FRI3	4.000	0.0549	-0.2446	0.0822
FRI5	4.000	0.0553	-0.2411	0.0827
FRI10	4.000	0.0562	-0.2424	0.0854
FNNI	4.001	0.0558	-0.2333	0.0809
MRNNI1	4.007	0.0567	-0.2107	0.0714
MRNNI3	4.008	0.0564	-0.2133	0.0701
MRNNI5	4.008	0.0562	-0.2130	0.0709
MRNNI10	4.009	0.0570	-0.2163	0.0728
FRNNI1	3.999	0.0543	-0.2385	0.0818
FRNNI3	4.000	0.0542	-0.2386	0.0814
FRNNI5	4.000	0.0544	-0.2387	0.0816

The data available have $n = 391$ observations. In simulations, a sample of 100 individuals were selected. Some of these individuals were deleted and it was pretended that they were missing. The respondents (for y) are selected in three different ways with response rate $p = 0.65$.

1. MCAR Missing completely at random: a uniform response mechanism.

2. MAR Missing at random: the missing values are randomly drawn by without replacement sampling within each class. The classes are formed using the x values. As financial stress increases, the chance of being missing increases.

3. NMAR Not missing at random: the missing values are randomly drawn by without replacement sampling within each class. The classes are formed using the y values. As self esteem decreases, the chance of being missing increases.

When the data are missing completely at random (MCAR), results of 10,000 simulations (sample size 100, 65 observed cases)

are presented in Table 4. The estimator of the mean is unbiased using all methods. Nearest neighbor methods (NN, FNNI, , MRNNI, FRNNI) tend to depress the magnitude of the slope estimate. FRNNI, which used without replacement sampling, had less bias than MRNNI, which used with replacement sampling.

When the data are missing at random (MAR), results of 10,000 simulations (sample size 100, 65 observed cases) are presented in Table 5. The MAR mechanisms was implemented by random selecting 48 (71%) of the smallest 68 x-values to be observed, but only 17 (53%) of the largest 32 to be observed. This is a relatively mild case of MAR. This choice of mechanism was made because 68% of the probability for a chisquare random variable with one degree of freedom is below its mean, and future simulations will investigate MAR cases for this distribution. Results are generally the same as before, but the methods CC and NN produce bias in point estimator of the mean of QFSE (Y). The other methods impute missing y-values using, in one way or another, based on the regression model and correct this bias. Further discussion of this phenomenon can be found in Little and Rubin (2002). The standard deviations of regression slope point estimates are larger than their counterparts in Table 4.

Table 6 presents results when data are missing not at random (NMAR). The NMAR mechanisms was implemented by random selecting 48 (71%) of the *largest* 68 y-values to be observed, but only 17 (53%) of the *smallest* 32 to be observed. Thus, small values of self-esteem are less likely to be observed. This is a relatively mild case of NMAR. All methods exhibit bias for the mean of Y and for the regresion slope. Standard deviations of point estimates are smaller than their MCAR and MAR counterparts. Nearest neighbor methods still exhibit more bias for the slope than other methods.

7. Summary and Discussion

Fractional regression nearest neighbor imputation was defined and studied through simulation. It was found that the distribution of the predictor values can have an effect on the performance of this and other nearest neighbor algorithms. Future work will investigate modifications of nearest neighbor matching algorithms to address this issue. Future simulations will consider additional populations and population regression models.

Methods were applied to data from the Iowa Family Transitions Project. It was demonstrated that the when the data are missing not at random (NMAR) that the missing data methods considered do not remove bias in the estimate of a mean or a regression slope. Future work will consider nearest neighbor methods to adjust for a suspected bias.

Future work also will consider variance estimation and the coverage of confidence intervals based on the methods in this paper.

Table 5: Monte Carlo means and standard deviations of point estimators of the mean of the outcome variable and of the linear regression slope based on 10000 replications. Variable X is QFFINSTR, Y is QFSE, $n = 100$, and the missing data mechanism is MAR.

	Mean of Y		Slope	
Method	Mean	SD	Mean	SD
Full data	4.000		-0.2410	
CC	4.015	0.0541	-0.2497	0.0861
NN	4.008	0.0607	-0.2031	0.0856
SR1	3.997	0.0603	-0.2495	0.0939
SR2	3.997	0.0600	-0.2497	0.0933
MI	3.997	0.0634	-0.2495	0.0900
FRI1	3.997	0.0553	-0.2498	0.0874
FRI3	3.997	0.0557	-0.2508	0.0868
FRI5	3.999	0.0562	-0.2432	0.0880
FRI10	3.999	0.0570	-0.2445	0.0910
FNNI	4.000	0.0568	-0.2318	0.0847
MRNNI1	4.011	0.0570	-0.2095	0.07315
MRNNI3	4.012	0.0570	-0.2107	0.0715
MRNNI5	4.013	0.0568	-0.2078	0.0729
MRNNI10	4.013	0.0571	-0.2105	0.0749
FRNNI1	3.996	0.0545	-0.2296	0.0800
FRNNI3	4.001	0.0548	-0.2380	0.0787
FRNNI5	3.998	0.0565	-0.2407	0.0850

Table 6: Monte Carlo means and standard deviations of point estimators of the mean of the outcome variable and of the linear regression slope based on 10000 replications. Variable X is QFFINSTR, Y is QFSE, $n = 100$, and the missing data mechanism is NMAR.

	Mean of Y		Slope	
Method	Mean	SD	Mean	SD
Full data	4.000		-0.2410	
CC	4.044	0.0485	-0.2321	0.0815
NN	4.042	0.0554	-0.1949	0.0809
SR1	4.036	0.0555	-0.2322	0.0890
SR2	4.036	0.0549	-0.2324	0.0886
MI	4.036	0.0579	-0.2322	0.0840
FRI1	4.036	0.0504	-0.2321	0.0829
FRI3	4.036	0.0505	-0.2336	0.0817
FRI5	4.037	0.0508	-0.2285	0.0826
FRI10	4.037	0.0516	-0.2296	0.0849
FNNI	4.039	0.0510	-0.2184	0.0797
MRNNI1	4.048	0.0514	-0.1997	0.0705
MRNNI3	4.049	0.0504	-0.2001	0.0690
MRNNI5	4.049	0.0509	-0.1998	0.0695
MRNNI10	4.049	0.0516	-0.2010	0.0715
FRNNI1	3.971	0.0533	-0.2337	0.0811
FRNNI3	3.973	0.0532	-0.2407	0.0801
FRNNI5	3.973	0.0536	-0.2357	0.0803

Acknowledgements

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. The authors thank Dr. Fred Lorenz (ISU, Statistics and Sociology) for use of the data set and Dr. Shinsoo Kang (Korea) for discussions.

References

Cheng, J.H., and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, 16(2): 113-131.

Chen, J.H., and Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association*, 96(453): 260-269.

Conger, R.D., Elder, G.H. Jr., Lorenz, F.O., Conger, K.J., Simons, R.L., Whitbeck, L.B., Huck, S., and Melby, J. (1990). Linking economic hardship to marital quality and instability. *Journal of Marriage and the Family*, 52, 643-656.

Conger, R.D., Lorenz, F.O., and Wickrama, K.A.S. (2004). *Continuity and change in family relations: Theory, methods, and empirical findings.* Mahwah, NJ: Erlbaum.

Durrant, G.B. (2005). Imputation methods for handling item-nonresponse in the social sciences: A methodological review. *NCRM Methods Review Papers*, NCRM/002, ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute (S3RI), University of Southampton.

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

Fuller, W.A., and Kim, J.K. (2005b). Replicated nearest neighbor imputation. Presentation at *IASS 55: A Satellite Conference of ISI 55: Complex sampling, retrospective sampling and missing data*. Auckland, New Zealand, September 2, 2005.

Kalton, G. (1983). *Compensating for missing survey data*. Institute of Social Research (Ann Arbor).

Kalton, G., and Kasprzyk, D. (1982). Imputing for missing survey responses. *ASA Proc. of the Section on Survey Research Methods*, American Statistical Association (Alexandria, VA), 22-31.

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

Kim, J.K. (2003). Fractional imputation using regression imputation model. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 2145-2152.

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

Laaksonen, S. (2000). Regression-based nearest neighbour hot decking. *Computational Statistics*, 15, 65-71.

Lessler, J.T., and Kalsbeek, W.D. (1992). *Nonsampling error in surveys*. John Wiley & Sons (New York; Chichester). Section 8.2.

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6, 287-297.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical analysis with missing data*. John Wiley & Sons (New York; Chichester).

Lorenz, F.O., Conger, R.D., Simons, R.L., Whitbeck, L.B., and Elder, G.H. Jr. (1991). Economic pressure and marital quality: An illustration of the method variance problem in the causal modeling of family processes. *Journal of Marriage and the Family*, 53, 375-388.

Nordholt, E.S. (1998). Imputation: Methods, simulation experiments and practical examples. *International Statistical Review*, 66(2): 157-180.

Rancourt, E. (1999) Estimation with nearest neighbor imputation at Statistics Canada. *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association (Alexandria, VA), 131-138.

Rancourt, E., Särndal, C., and Lee, H. (1994) Estimation of the variance in the presence of nearest neighbour imputation. *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association (Alexandria, VA), 888-893.

Rao, J.N.K. (1996). On variance estimation with imputed survey data (Pkg: p473-520). *Journal of the American Statistical Association*, 91, 499-506.

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

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*. John Wiley & Sons (New York; Chichester).

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Wickrama, K.A.S., Conger, R.D., and Lorenz, F.O. (1995). Work, family, lifestyle and health. *Journal of Behavioral Medicine*, 18, 97 - 111.