Journal of Statistical Software

February 2009, Volume 29, Issue 9. *http://www.jstatsoft.org/*

MIDAS: A SAS Macro for Multiple Imputation Using Distance-Aided Selection of Donors

Juned Siddique
Northwestern University

Ofer Harel
University of Connecticut

Abstract

In this paper we describe **MIDAS**: a SAS macro for multiple imputation using distance-aided selection of donors which implements an iterative predictive mean matching hot-deck for imputing missing data. This is a flexible multiple imputation approach that can handle data in a variety of formats: continuous, ordinal, and scaled. Because the imputation models are implicit, it is not necessary to specify a parametric distribution for each variable to be imputed. **MIDAS** also allows the user to address the sensitivity of their inferences to different assumptions concerning the missing data mechanism. An example using **MIDAS** to impute missing data is presented and **MIDAS** is compared to existing missing data software.

Keywords: hot-deck, missing data, predictive mean matching, approximate bayesian bootstrap, abb, not missing at random, nmar, nonignorable.

1. Introduction

Missing values are a problem in many data sets and are ubiquitous in the social and health sciences. A common and practical method for dealing with missing data is multiple imputation where missing values are replaced with two or more plausible values. Multiple imputation techniques that specify an explicit Bayesian model have desirable theoretical properties that lead to inferences that are valid when the model does a good job of representing available information (Rubin 1987). However, many such techniques can be difficult to implement in that they are often tailored to specific data types and require the imputer to model the joint distribution of the data, which is hard to do in high-dimensional problems with many variables. Moreover, most multiple imputation procedures assume that the missing data are *ignorable* as defined by Rubin (1976) where the probability of missingness depends only on observed values. This assumption is questionable in many applications, and even when it is a reasonable assumption, it is important for the analyst to check how sensitive inferences are to different assumptions concerning the missing data mechanism.

We describe a SAS (SAS Institute Inc. 2003) macro **MIDAS**: multiple imputation using distance-aided selection of donors, that implements the methods described in Siddique and Belin (2008a,b). These methods, iterative hot-deck multiple imputation with distance-based donor selection, can handle a variety of data types yet still incorporate desirable features of Bayesian approaches such as the ability to reflect parameter uncertainty, handle missing covariate values, and incorporate all available information into the imputation model. In addition, the methods allow the user to impute nonignorable missing data.

1.1. Background

The properties of missing data methods may depend strongly on the mechanism that led to the missing data. A particularly important question is whether the fact that variables are missing is related to the underlying values of the variables in the data set (Little and Rubin 2002).

Specifically, Rubin (1976) classifies the reasons for missing data as either *ignorable* or *nonignorable*. In a data set where variables X are fully observed and variables Y have missing values, the missingness in Y is deemed ignorable if the missing Y values are only randomly different from observed Y values when conditioning on the X values. Nonignorable missingness asserts that even though two observations on Y (one observed, one missing) have the same X values, their Y values are systematically different. Rubin and Schenker (1991) give an example where the missing Y are typically 20 percent larger than observed Y for the same values of X. The role of nonignorability assumptions has been discussed in the context of a variety of applied settings; see e.g., Little and Rubin (2002, Chapter 15), Belin *et al.* (1993), Wachter (1993), Rubin *et al.* (1995), Schafer and Graham (2002), Demirtas and Schafer (2003).

Imputation is a common and practical method for dealing with missing data where missing values are replaced with plausible values. Simply imputing missing values once, and then proceeding to analyze a data set as if there never were any missing values (or as if the imputed values were the observed values) fails to account for the uncertainty due to the fact that the analyst does not know the values that might have been observed. No matter how successful an imputation procedure has been in eliminating nonresponse bias, it is important to account for this additional uncertainty.

Rubin (1987) proposed handling the uncertainty due to missingness through the use of multiple imputation. Multiple imputation refers to the procedure of replacing each missing value with $D \geq 2$ imputed values. Then D imputed data sets are created, each of which can be analyzed using complete data methods. Using rules that combine within-imputation and between-imputation variability (Rubin 1987), inferences are combined across the D imputed data sets to form one inference that properly reflects uncertainty due to nonresponse under that model. However, creating multiple imputations and combining complete-data estimates does not insure that the resulting inferences will be valid. Rubin (1987) defines the conditions that a multiple imputation procedure must meet in order to produce valid inferences and be deemed a *proper* multiple imputation procedure. To satisfy these conditions, a multiple imputation procedure must provide randomization-valid inferences in the complete data and must represent both the sampling uncertainty in the imputed values and the estimation uncertainty associated with either explicit or implicit unknown parameters.

A hot-deck is an imputation method where missing values (donees) are replaced with observed values from donors deemed exchangeable with the donees. There are many benefits to hot-deck imputation including: 1) imputations tend to be realistic since they are based on values observed elsewhere; 2) imputations will not be outside the range of possible values; and 3) it is not necessary to define an explicit model for the distribution of the missing values. Because of the simplicity of the hot-deck approach and these desirable properties, it is a popular method of imputation, especially in large sample survey settings where there is a large pool of donors.

1.2. Research overview

The outline for the rest of this paper is as follows. In Section 2, we briefly describe the hot-deck imputation methods of Siddique and Belin (2008a,b) that are implemented in **MIDAS**. Section 3 provides documentation on using **MIDAS** and the macro inputs. Section 4 gives an example of using **MIDAS** to impute and analyze a data set with missing values, first assuming that the missing data mechanism is ignorable, then assuming that it is not. Section 5 compares **MIDAS** to eight other well-known missing data software programs that were recently evaluated in Horton and Kleinman (2007). Section 6 offers practical guidelines for multiply imputing missing data using **MIDAS**. Section 7 gives concluding remarks and further thoughts on implementing **MIDAS**.

2. Methods

In this section we briefly describe the predictive mean matching hot-deck imputation method of Siddique and Belin (2008a) and the nonignorable approximate Bayesian bootstrap method of Siddique and Belin (2008b) that are implemented in **MIDAS**.

2.1. Hot-deck multiple imputation using distance-based donor selection

Since hot-deck procedures are most tractable when imputing one variable at a time, for the remainder of this paper, define Y to be a single variable with missing values. Y_{obs} consists of the values of Y that are observed and Y_{mis} consists of the values of Y that are missing. Let n_{obs} and n_{mis} be the number of cases associated with Y_{obs} and Y_{mis} respectively. In predictive mean matching (Little 1988; Schenker and Taylor 1996), values Y_{obs} are regressed on a set of observed variables, say X. Then, using the regression parameters calculated on the observed data, predicted values $\hat{Y}$ are calculated for all Y. Finally, Y_{mis} values are imputed using Y_{obs} values whose predicted $\hat{Y}$ values are similar. An *approximate Bayesian bootstrap* (ABB) (Rubin and Schenker 1986; Demirtas *et al.* 2007) is a method for incorporating parameter uncertainty into hot-deck imputation models. An ignorable ABB first draws n_{obs} cases randomly with replacement from Y_{obs} to create Y^*_{obs}. Donors for imputing missing values are then selected from this new set of "observed" cases. For multiple imputation, D bootstrap samples are drawn so that the imputed values are drawn from D different sets of donors.

Siddique and Belin (2008a) describe a distance-based donor selection approach with an ABB where donors are selected with probability inversely proportional to their distance from the donee. Using only those rows (cases) where Y is observed, the ABB is performed by drawing n_{obs} rows with replacement. Let $w_j, j = 1, \ldots, n_{\text{obs}}$ designate the number of times the row belonging to $y_j \in Y_{\text{obs}}$ was chosen with replacement in the ABB. Let W represent a diagonal matrix of w_j values. The formula for calculating the predicted values is $\hat{Y} = X\hat{B}$ where $\hat{B} = (X^\top WX)^{-1}X^\top WY_{\text{obs}}$. For a given donee, let D_{0i}^k be the distance between donee 0 and donor i, where the distance is the absolute difference in predicted values raised to a power k with a non-zero offset to avoid complexities posed by zero distances. That is,

$$D_{0i}^k = \left(|\hat{y}_0 - \hat{y}_i| + \delta\right)^k, \tag{1}$$

where

$$\delta = \min|\hat{y}_0 - \hat{y}_i| \text{ for all } \; i = 1, \ldots, n_{\text{obs}} \text{ where } \; \hat{y}_0 \neq \hat{y}_i.$$

In settings where no two cases i and j have $\hat{y}_i = \hat{y}_j$ (i.e., when no two individuals have the same pattern of observed covariates or identical predicted values), then the δ offset term is dropped (although retaining it in the procedure is apt to have little practical consequence).

Using the distance defined in Equation 1, the donor i selection probability $l_i^k(\hat{y}_0)$ for donee 0 is

$$l_i^k(\hat{y}_0) = \frac{\frac{1}{D_{0i}^k} w_i}{\sum_{j=1}^{n_{\text{obs}}} \frac{1}{D_{0j}^k} w_j}. \tag{2}$$

Equation 2 ensures that

$$\sum_{i=1}^{n_{\text{obs}}} l_i^k(\hat{y}_0) = 1$$

and that given observed donor values $Y_{\text{donor}} = (Y_1, \ldots, Y_{n_{\text{obs}}})$ and selection probabilities $l_{\text{donor}}^k(\hat{y}_0) = (l_1^k(\hat{y}_0), \ldots, l_{n_{\text{obs}}}^k(\hat{y}_0))$, the expected value of the imputation for donee 0 is

$$E(y_0|Y_{\text{donor}}, l_{\text{donor}}^k(\hat{y}_0)) = \sum_{i=1}^{n_{\text{obs}}} l_i^k(\hat{y}_0)Y_i.$$

With this approach, donors closest to the donee have greatest probability of selection, but all donors are eligible and have some non-zero selection probability. The exponent k in Equation 1 is a *closeness* parameter, which adjusts the probability of selection assigned to the closest donors. As $k \to \infty$ this procedure amounts to a nearest-neighbor hot-deck where the donor whose predicted mean is closest to the donee is always chosen. Conversely, when k equals 0, each donor has equal probability of selection, which is equivalent to a simple random hot-deck. In practice, a value of k somewhere between these two extremes is chosen by the imputer. Siddique and Belin (2008a) considered an example where a closeness parameter value around 3 appeared to be reasonable to favor nearby donors while allowing donor probabilities to decline smoothly as a function of distance.

In addition to adjusting the probability of selection assigned to the closest donors, the closeness parameter also has an impact on the bias and variance of the imputed values. Small values of the closeness parameter imply lower variance (averaging over more donors) but presumably higher bias (since it may not be plausible to assume that all donors are equally good matches).

When covariates in the predictive mean matching models have missing values, starting values are introduced. Then, once all variables have been imputed once, they are re-imputed, this time replacing starting values with imputed values. This procedure is iterated until convergence diagnostics are achieved (Siddique and Belin 2008a). Siddique and Belin (2008a) showed in one setting that 10 iterations of the hot-deck procedure produced estimates that were not significantly different from estimates resulting from iterating the procedure until more formal convergence diagnostics were satisfied.

Multiple imputation is incorporated into the method to reflect the uncertainty of the imputations by performing the procedure D times to create D complete data sets. Each data set is analyzed separately, and inferences are combined using the rules described by Rubin (1987).

2.2. Implementing a nonignorable approximate Bayesian bootstrap

Rubin and Schenker (1991) discuss how an ABB can be modified to handle nonignorable missing data. Instead of drawing n_{obs} cases of Y_{obs} randomly with replacement (i.e., with equal probability), they suggest drawing n_{obs} cases of Y_{obs} with probability proportional to Y^c_{obs} so that the probability of selection for for $y_i \in Y_{\text{obs}}$ is

$$\frac{y_i^c}{\sum_{j=1}^{n_{\text{obs}}} y_j^c}. \tag{3}$$

This skews the nonrespondents to have typically larger (when $c > 0$ and $y_j > 0$) values of Y than respondents. Siddique and Belin (2008b) refer to the ABBs where values of Y_{obs} are drawn with probability proportional to Y^c_{obs} where $c = -1, 1, 2,$ and 3, as an "inverse-to-size ABB", "proportional-to-size ABB", "proportional-to-size-squared ABB", and "proportional-to-size-cubed ABB" respectively.

In addition to these nonignorable ABBs proposed by Rubin and Schenker (1991) where n_{obs} cases of Y_{obs} are drawn with probability proportional to Y^c_{obs}, Siddique and Belin (2008b) consider a number of variations on this idea. Siddique and Belin (2008b) describe a nonignorable ABB where n_{obs} cases of Y_{obs} are drawn with probability proportional to $[|Y_{\text{obs}} - Q_p(Y_{\text{obs}})|]^c$ where the notation $Q_p(Y_{\text{obs}})$ represents the p-th quantile of Y_{obs}. For example, when $p = 2$ so that $Q_p(Y_{\text{obs}})$ is the median of the observed Y values, the implication of drawing with probability proportional to the distance from the median is to favor values for the non-respondents with either larger *or* smaller values than respondents with the same set of covariates (when $c > 0$). They refer to this approach as a "U-shaped ABB" because observations in the extremes of the distribution of Y_{obs} have greater weight than observations in between that are close to the median.

In a variation of this idea with $p = 1$, Siddique and Belin (2008b) refer to the ABB that centers the donor sizes around the 1st quantile as a "fishhook ABB", because this ABB mostly favors large values but retains a U-shaped pattern featuring a slight upturn in the weight given to the smallest observed values.

For all nonignorable ABBs described above, when values of Y_{obs} are less than or equal to 0, the values of Y_{obs} need to be transformed to ensure that the selection probabilities in the nonignorable ABB are positive and (in the case where Y_{obs} are drawn with probability proportional to Y^c_{obs}, $c > 0$) that the selection probability for $y_i \in Y_{\text{obs}}$ is greater than the selection probability for $y_j \in Y_{\text{obs}}$ when $y_i > y_j$. Define α and β to be the smallest and second smallest values of Y_{obs} respectively where $\alpha \neq \beta$. Transform $y_i \in Y_{\text{obs}}$ using $y_i + |\alpha| + |\alpha - \beta|$. Then Equation 3 is rewritten as

$$\frac{(y_i + |\alpha| + |\alpha - \beta|)^c}{\sum_{j=1}^{n_{\text{obs}}} (y_j + |\alpha| + |\alpha - \beta|)^c}.$$

This transformation is used only for calculating the selection probabilities. The original values of Y_{obs} are used for imputation.

3. Using MIDAS

MIDAS is written as a SAS macro using the Base SAS, SAS/STAT, and SAS/IML modules. The macro is called as:

```
%MIDAS(y, dataset, key, covar, itnum = 1, close = 3, step = NONE,
  seed = 0, start = MEAN, abbtype = NONE, pps = 0);
```

The arguments are defined below.

`y` Target variable to be imputed.

`dataset` Data set name.

`key` A variable that uniquely identifies each observation (row) in the dataset (e.g., subject ID). Must be numeric.

`covar` List of covariates to be used in the imputation model. Can have missing values for the first iteration but not for subsequent iterations. If there are missing covariates and the iteration is greater than 1 then **MIDAS** will fail to produce imputations.

`itnum` Iteration number. Intended to be part of a loop. Default is `1`.

`close` Closeness parameter as described in Siddique and Belin (2008a). A closeness parameter equal to `0` is equivalent to a simple random hot-deck where each donor has equal probability of selection. As the closeness parameter approaches infinity, the hot-deck becomes a nearest neighbor hot-deck where the closest donor is always chosen. Default is `3`.

`step` Variable selection procedure for choosing variables in the predictive mean matching model. Choices are `NONE`, `BACKWARD`, and `FORWARD`. Default is `NONE`.

`seed` Random number seed. Default is `0`.

`start` Type of imputation procedure (mean imputation or simple random imputation) to get starting values for missing covariates. Inputs are `MEAN` or `SRS`. Mean imputation imputes missing covariate values with the mean of the observed values. Simple random imputation imputes each missing covariate value by randomly drawing from the observed values. The default is `MEAN`. Missing covariates are only filled in for the first iteration.

`abbtype` Center observations for ABB or not (`NONE`, `MEAN`, `MEDIAN`, `Q1`, `Q3`). Default is `NONE`.

`pps` For ABB. Choose sample with probability proportional to y^{pps}. When `pps = 0`, choose among observed with equal probability. Default is `0`.

For each variable that has been imputed, **MIDAS** creates a new variable with the prefix `i_`, where `i_varname` is equal to 1 if the corresponding value of `varname` has been imputed, and 0 otherwise. The ABB is implemented during the first iteration and a bootstrap weight variable is created with the suffix `bwt`, where `varnamebwt` indicates the number of times the observed value of `varname` was selected with replacement in the ABB. Starting values are only created during the first iteration. Variables that appear as covariates in imputation models should be imputed so that there are no covariates with missing values after the first iteration. After the first iteration, previously imputed values of the target variable are deleted then re-imputed.

4. Data example: The St. Louis Risk Research Project

The St. Louis Risk Research Project (SLRRP) was an observational study to assess the effects of parental psychological disorders on various aspects of child development. In a preliminary cross-sectional study, data were collected on 69 families having two children each. The families were classified into three risk groups for parental psychological disorders. The children were classified into two groups according to the number of adverse psychological symptoms they exhibited. Standardized reading and verbal comprehension scores were also collected for the children. Each family is thus described by four continuous and three categorical variables. Rates of missingness range from 0% to 43% (see Table 1).

Because of its mixture of continuous and categorical variables with missing values, the SLRRP data set has become a classic data set for evaluating imputation methods for mixed data types. See Little and Schluchter (1985), Schafer (1997), Liu and Rubin (1998), Raghunathan *et al.* (2001).

Using the SLRRP data, Raghunathan *et al.* (2001) investigated the impact of parental psychological disorders on childhood reading and verbal scores after adjusting for the number of symptoms. Raghunathan *et al.* (2001) used scores on the log scale, resulting in the following mixed-effects regression model:

$$\log R_{ic} = \alpha_0 + \alpha_1 G_{1i} + \alpha_2 G_{2i} + \alpha_3 D_{ic} + \delta_i + \epsilon_{ic}$$

where R_{ic} is the reading score for child c in family i; $G_{1i} = 1$ if family i is classified as a moderate risk group and 0 otherwise; $G_{2i} = 1$ if family i is classified as a high risk group and 0 otherwise; D_{ic} is the symptom level for child c in family i; and δ_i are random effects to account for interclass correlation between two children within the same family. The terms δ_i and ϵ_{ic} are assumed to be mutually independent normal random variables with mean 0 and variances σ^2_δ and σ^2_ϵ respectively.

Variable	Levels	Code	Percent missing
Parental risk group	1 = low, 2 = moderate, 3 = high	G	0
Symptoms, child 1	0 = low, 1 = high	D_1	41
Symptoms, child 2	0 = low, 1 = high	D_2	41
Reading score, child 1	continuous	R_1	30
Verbal score, child 1	continuous	V_1	43
Reading score, child 2	continuous	R_2	23
Verbal score, child 2	continuous	V_2	25

Table 1: Variables from the SLRRP.

Below, we provide **MIDAS** code for imputing the SLRRP data and SAS code for analyzing the data. The imputation is nested within a macro to facilitate the use of iteration and the creation of 5 multiply imputed data sets. For each of the 5 imputed data sets, 10 iterations are performed to reduce dependence on starting values and imputation order. The seed is changed before imputing each variable to incorporate additional uncertainty into the imputations. Variables are imputed one at a time. Each variable to be imputed uses all other variables in the data set in its imputation model. After the first iteration, missing covariate values are replaced with imputed values instead of starting values and previously imputed values for target variables are deleted and re-imputed. The default closeness parameter value of 3 is used, as are default starting values based on mean imputation. No stepwise procedures are incorporated into the predicted mean matching regression models. In the next section we demonstrate the use of a nonignorable ABB. The results are presented in Table 2.

```
%include '<full path>\MIDAS.sas';

%macro slrrp_macro(mult,itnum);

%do k=1 %to &mult; *do across multiple data sets;

data slrrp_&k;
set slrrp;
_Imputation_=&k;
run;

%do its=1 %to &itnum; *do across multiple iterations;
*The seed must be different for each variable and data set;
%let seed1=%eval(9999+&k+&its);

*Impute Child 1 Symptoms;
%let seed1=%eval(&seed1+1);
%MIDAS(D1, slrrp_&k, mid, D2 R1 R2 V1 V2 G2 G3, itnum=&its, seed=&seed1);

*Impute Child 2 Symptoms;
%let seed1=%eval(&seed1+1);
%MIDAS(D2, slrrp_&k, mid, D1 R1 R2 V1 V2 G2 G3, itnum=&its, seed=&seed1);

*Impute Child 1 Reading Score;
%let seed1=%eval(&seed1+1);
%MIDAS(R1, slrrp_&k, mid, R2 V1 V2 D1 D2 G2 G3, itnum=&its, seed=&seed1);

*Impute Child 2 Reading Score;
%let seed1=%eval(&seed1+1);
%MIDAS(R2, slrrp_&k, mid, R1 V1 V2 D1 D2 G2 G3, itnum=&its, seed=&seed1);

*Impute Child 1 Verbal Score;
%let seed1=%eval(&seed1+1);
%MIDAS(V1, slrrp_&k, mid, V2 R1 R2 D1 D2 G2 G3, itnum=&its, seed=&seed1);
```

```
*Impute Child 2 Verbal Score;
%let seed1=%eval(&seed1+1);
%MIDAS(V2, slrrp_&k, mid, V1 R1 R2 D1 D2 G2 G3, itnum=&its, seed=&seed1);

%end; *end iteration loop;

*append multiply imputed data sets;
proc datasets library=work nolist;
append base=work.slrrp_impute data=work.slrrp_&k;
run;
quit;

%end; *end multiple data sets loop;

%mend slrrp_macro;

%slrrp_macro(5,10);
```

Then, to perform the analysis using PROC MIXED in SAS, we must convert the data set from a horizontal format to a vertical format.

```
data slrrp_vert;
set slrrp_impute;
family=mid;
logR1=log(R1);
logR2=log(R2);
child=1; symptoms=D1; logread=logR1; output;
child=2; symptoms=D2; logread=logR2; output;
run;
```

The random intercept model for log reading score is fit using the SAS procedure PROC MIXED separately by imputed data set. Parameter estimates are exported to a new data set.

```
proc mixed data=slrrp_vert;
model logread = G2 G3 symptoms/solution;
random intercept/sub=family type=un g gcorr;
by _Imputation_;
ods output SolutionF=mixparms1;
title 'Dependent Variable Log Reading Score';
run;
quit;
```

The SAS procedure PROC MIANALYZE is used to combine parameter estimates from imputed data sets using the rules described by Rubin (1987).

```
proc mianalyze parms=mixparms1;
modeleffects Intercept G2 G3 symptoms;
run;
```

4.1. Nonignorable missing data

The above imputation method assumed that the missing data were ignorable as defined by Rubin (1976) where the probability of missingness depends only on observed values. However, this assumption is unlikely in most applications and even when it is a reasonable assumption, it is important for the analyst to check how sensitive inferences are to different assumptions concerning the missing data mechanism. In this section, we impute the SLRRP data again, this time using a different nonignorable ABB for each imputed data set which Siddique and Belin (2008b) refer to as a *mixture ABB*. Here, we assume that missing reading and verbal scores tend to be lower than observed values with the same covariates. The mixture ABB approach used here is an inverse-to-size-cubed ABB, inverse-to-size-squared ABB, inverse-to-size ABB, ignorable ABB, and proportional-to-size ABB. The code is the same as above except that we add the input pps in the macro statements for reading score and verbal score. As recommended by Siddique and Belin (2008b), we use a closeness parameter value of 1 since larger closeness parameter values can reduce the effectiveness of the nonignorable ABB.

```
%include '<full path>\MIDAS.sas';

%macro slrrp_macro(mult,itnum);

%do k=1 %to &mult; *do across multiple datasets;

*Use a different ABB for each imputed data set;
%let abbvalue=%eval(&k-4);

data slrrp_&k;
set slrrp;
_Imputation_=&k;
run;

%do its=1 %to &itnum; *do across multiple iterations;

*The seed must be different for multiple datasets;
%let seed1=%eval(9999+&k+&its);

*Impute Child 1 Symptoms;
%let seed1=%eval(&seed1+1);
%MIDAS(D1, slrrp_&k, mid, D2 R1 R2 V1 V2 G2 G3, itnum=&its, seed=&seed1);

*Impute Child 2 Symptoms;
%let seed1=%eval(&seed1+1);
%MIDAS(D2, slrrp_&k, mid, D1 R1 R2 V1 V2 G2 G3, itnum=&its, seed=&seed1);

*Impute Child 1 Reading Score using a nonignorable ABB and
a closeness parameter value equal to 1;
%let seed1=%eval(&seed1+1);
```

```
%MIDAS(R1, slrrp_&k, mid, R2 V1 V2 D1 D2 G2 G3, itnum=&its,
close=1, seed=&seed1, pps=&abbvalue);

*Impute Child 2 Reading Score using a nonignorable ABB and
a closeness parameter value equal to 1;
%let seed1=%eval(&seed1+1);
%MIDAS(R2, slrrp_&k, mid, R1 V1 V2 D1 D2 G2 G3, itnum=&its,
close=1, seed=&seed1, pps=&abbvalue);

*Impute Child 1 Verbal Score using a nonignorable ABB and
a closeness parameter value equal to 1;
%let seed1=%eval(&seed1+1);
%MIDAS(V1, slrrp_&k, mid, V2 R1 R2 D1 D2 G2 G3, itnum=&its,
close=1, seed=&seed1, pps=&abbvalue);

*Impute Child 2 Verbal Score using a nonignorable ABB and
a closeness parameter value equal to 1;
%let seed1=%eval(&seed1+1);
%MIDAS(V2, slrrp_&k, mid, V1 R1 R2 D1 D2 G2 G3, itnum=&its,
close=1, seed=&seed1, pps=&abbvalue);

%end; *end iteration loop;

*append multiply imputed data sets;
proc datasets library=work nolist;
append base=work.slrrp_impute data=work.slrrp_&k;
run;
quit;

%end; *end multiple datasets loop;

%mend slrrp_macro;

%slrrp_macro(5,10);
```

Table 2 displays the mixed-effects model regression coefficients with standard errors and intra-class correlations from the regressions for log reading score from a complete-case analysis, **MIDAS** imputation assuming ignorability, and **MIDAS** imputation assuming nonignorability. In all three analyses, reading scores in the moderate and high parental risk groups are lower than in the low parental risk group. These differences are less pronounced for reading scores under the nonignorable imputation model where we assumed that the missing reading and verbal scores were lower than observed reading and verbal scores conditional on other covariates. The intra-class correlation coefficient is largest for the complete-case analysis and smallest when nonignorability is assumed. These results may reflect the additional uncertainty multiple imputation procedures can incorporate into parameter estimates.

Parameter	Complete case	Assume ignorable	Assume nonignorable
Intercept	4.744 (0.038)***	4.686 (0.027)***	4.694 (0.034)***
Moderate risk	−0.139 (0.053)*	−0.110 (0.040)**	−0.099 (0.042)*
High risk	−0.182 (0.056)**	−0.117 (0.045)*	−0.106 (0.042)*
Symptoms	0.025 (0.043)	0.060 (0.038)	0.030 (0.029)
Intra-class correlation	0.337	0.316	0.263

Table 2: Mixed-effects model estimates of regression coefficients with standard errors and intra-class correlations for SLRRP reading scores using complete cases, **MIDAS** imputation assuming ignorablity, and **MIDAS** imputation assuming nonignorability. p values are given for regression coefficients only: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5. Comparing MIDAS with other missing data software

Horton and Kleinman (2007) compared eight commonly used missing data software packages using the Kids' Inpatient Database (KID) for the year 2000 provided by the Healthcare Cost and Utilization Project – HCUP (2003). The eight software packages were **Amelia II** (Honaker *et al.* 2008), **Hmisc** (Harrell Jr 2008), **ice** (Royston 2005), **IVEware** (Raghunathan *et al.* 2002), **LogXact** (Cytel Inc. 2006), **mice** (van Buuren and Oudshoorn 2007), SAS PROC MI (SAS Institute Inc. 2003), and the S-PLUS **Missing** library (Insightful Corp. 2003). The KID data set, which is publicly available for a fee from the Agency for Healthcare Research and Quality, collects data from states on child hospitalization to improve the quality of health care. Horton and Kleinman (2007) investigated what factors predicted whether a pediatric subject with a psychiatric or substance abuse diagnosis had a routine discharge from the hospital.

The outcome in their model was routine discharge versus non-routine discharge. Predictors in the logistic regression included an indicator for gender, age, length of stay, admission type, admission season, admission on weekend, number of diagnoses on original record, race (white, black, hispanic, other), and total charges.

The data set consisted of 134,774 observations. There was a substantial amount of missingness in their data set. Admission type was missing for 11.2% of cases. Race was missing for 16.2%, total charges for 3.7%, and season was missing for 11.6%. A total of 79,574 (59%) observations had complete data. See Horton and Kleinman (2007) for more details regarding the data, the analysis model, and the patterns of missingness.

We imputed the KID data using **MIDAS** and then analyzed the data using the same logistic regression model as Horton and Kleinman. Each category of race, season, and admission type were converted into binary variables and imputed one at a time. Race and admission type were imputed in order from the least common to the most common category. Five imputations were made for each missing value with ten iterations, a closeness parameter value equal to 3, and mean imputation for starting values. The covariates in each imputation model were the same independent and dependent variables that appeared in the analysis model.

Table 3 has been reproduced from Horton and Kleinman (2007) and displays the logistic regression coefficients for weekend admission (WEEKEND), gender (FEMALE), and total charges (TOTCHG) from the KID analysis using eight different missing data software packages as well as the complete-case analysis. In the last line of the table, the results from using

Package	WEEKEND	FEMALE	TOTCHG
Complete case	−0.058 (0.026)	0.089 (0.021)	−0.004 (0.0010)
Amelia II	−0.027 (0.020)	0.103 (0.016)	−0.005 (0.0005)
Hmisc	−0.020 (0.020)	0.099 (0.016)	−0.005 (0.0005)
ice	−0.020 (0.020)	0.099 (0.016)	−0.004 (0.0005)
IVEware	−0.021 (0.020)	0.100 (0.016)	−0.004 (0.0005)
mice	−0.021 (0.020)	0.100 (0.016)	−0.004 (0.0005)
LogXact	−0.026 (0.020)	0.105 (0.016)	−0.005 (0.0005)
SAS `PROC MI`	−0.036 (0.021)	0.119 (0.017)	−0.003 (0.0006)
S-PLUS **Missing**	−0.018 (0.020)	0.098 (0.016)	−0.004 (0.0005)
MIDAS	−0.021 (0.020)	0.101 (0.016)	−0.004 (0.0005)

Table 3: Results (in terms of log OR and SE) for selected regression parameters for a variety of incomplete logistic regression models including **MIDAS**. The non-**MIDAS** results have been taken from Horton and Kleinman (2007, Table 4).

MIDAS have been added. **MIDAS** produces results very similar to the other packages. As Horton and Kleinman (2007) note, the parameter estimates for FEMALE and TOTCHG are similar for all missing data models relative to the complete-case estimator. The differing results for the WEEKEND parameter may indicate selection bias due to discarding those cases that are partially observed.

In addition to providing inferences similar to other imputation software packages, **MIDAS** has a number of features that are not available in any one of the packages in Table 3. **MIDAS** does not require the joint distribution of the data to be specified, no special accommodations are necessary to avoid imputing out of range or unrealistic values, and further analyses assuming that the missing data are nonignorable are possible using **MIDAS**.

6. Implementation guidelines

In this section we offer some practical guidelines for performing multiple imputation using **MIDAS**.

6.1. Selecting variables for the imputation model

In general, imputation models should use all available information to increase predictive power and to accommodate a large number of different data analyses (Meng 1994). Variables to be included in an imputation model should be associated either with the variable to be imputed or the probability that the variable is missing (Rubin 1976). When deciding which covariates should be included in an imputation model, we recommend the 'inclusive' imputation strategy of Collins *et al.* (2001) where in addition to the variables used in the analysis, the imputation model also incorporates 'auxiliary' variables that are used in the imputation procedure but are not incorporated in the analysis model. The simulation findings of Collins *et al.* (2001) show that there are noticeable gains in terms of increased efficiency and reduced bias from including auxiliary variables in addition to analysis variables.

Special care needs to be taken when imputing clustered data so that associations within clusters are preserved. When imputing longitudinal data, imputation models should condition on measurements made at prior and subsequent time points. When imputing hierarchical data like the SLRRP, conditioning on variables within the same cluster will not only preserve the correlation structure within the hierarchy, but may also provide additional predictive power. This was the approach taken when imputing the SLRRP data; when imputing a Child 1 variable, not only did we include all other Child 1 variables in our imputation model, we also included all Child 2 variables.

As Siddique and Belin (2008a) note, identifying uncongeniality (Meng, 1994) between imputation and analysis models is a challenge for hot-deck procedures because the imputation model is implicit. They also note that the implicit nature of the imputation model gives tremendous flexibility in incorporating a range of possible values so that the analysis procedure corresponds to the imputation procedure.

6.2. Implementing the nonignorable approximate Bayesian bootstrap

A unique feature of **MIDAS** is its ability to impute nonignorably missing data. When implementing the nonignorable ABB, we recommend using the mixture ABB approach of Siddique and Belin (2008b) where each imputed data set uses a different nonignorable (or ignorable) ABB. Siddique and Belin (2008b) showed that the mixture ABB appears to account for appropriate uncertainty and provide nominal coverage even when the missing data mechanism is nonignorable. Since large closeness parameter values reduce the effectiveness of the nonignorable ABB, Siddique and Belin (2008b) recommend using a closeness parameter value in the range of 1 to 2.

The mixture ABB procedure can be altered depending on what is perceived as the reason for missingness. For example, if the imputer believes that smaller values are more likely to be missing, but still wishes to incorporate uncertainty regarding nonignorability, then a mixture ABB can be chosen that is centered around an ABB that favors smaller donors. This was the strategy used to impute the reading and verbal scores in the SLRRP example, where inverse-to-size-cubed, inverse-to-size-squared, inverse-to-size, ignorable, and proportional-to-size ABBs were used.

A mixture ABB that is centered around an ignorable ABB (e.g., inverse-to-size-squared, inverse-to-size, ignorable, proportional-to-size, proportional-to-size-squared) is apt to provide inferences similar to an ignorable ABB, but with larger standard errors that presumably account for the uncertainly regarding the missing data mechanism. The U-shaped and fishhook ABBs can also be incorporated into a mixture ABB along with other ABBs if the imputer believes that these ABBs represent plausible missing data mechanisms.

Depending on the goals of the imputer, a nonignorable ABB can be used to provide a single inference that does not assume ignorability and/or to check the sensitivity of inferences to different ignorability assumptions. A desirable approach has been outlined by Daniels and Hogan (2008), namely 1) explore the sensitivity of inferences to unverifiable missing data assumptions, 2) characterize the uncertainty about these assumptions, and 3) incorporate subjective beliefs about the distribution of missing responses. We see connections between these goals and the use of nonignorable ABBs. Specifically, by analyzing data using several different ABBs, one can explore the sensitivity of inferences to different missing data assumptions. Use of the mixture ABB approach allows the analyst to characterize uncertainty about

ABB assumptions. And the choice of the ABB itself, whether it favors small or large donors incorporates subjective beliefs about why values are missing.

6.3. Multiple imputation diagnostics

Before running **MIDAS** we recommend that users confirm that the linear regression that will be used in each imputation model is estimable and that there are no problems with multicollinearity which can sometimes occur when one uses an inclusive imputation strategy.

We also suggest two diagnostics recommended by Abayomi *et al.* (2008) that compare observed and imputed values: density comparisons and bivariate scatterplots. These types of diagnostics can easily be done post-imputation using the `i_varname` indicator variable that **MIDAS** creates to identify observed and imputed values.

7. Concluding remarks

A major limitation to the software described here is computational time. Because each variable is imputed individually (rather than a model that specifies a joint distribution), **MIDAS** may require significant computational resources to impute a data set. The exact time required will depend on the number of imputed data sets, the number of iterations, the number of variables to be imputed, the number of observations in the data set, the amount of missingness in the data set, and the speed of the computer performing the imputations. However, in any data set with more than $10,000$ observations, **MIDAS** would probably exceed the amount of time most users would be willing to tolerate in a missing data procedure and a faster software package should be chosen.

Future versions of SAS will no doubt take advantage of multi-core processors to improve computing speed. It is possible that these advances will improve the performance of **MIDAS**. Otherwise, future releases of **MIDAS** will incorporate embedded C++ code to reduce computing time.

An additional limitation to **MIDAS** is that each variable to be imputed requires a separate macro statement. While this feature allows the user to specify a separate imputation model for each variable, a tradeoff is that in those situations where a user simply wants to specify a list of variables where each variable with missing values uses all other variables in its imputation model, the **MIDAS** notation can quickly become cumbersome.

One advantage of the distance-based hot-deck approach is that since we are using our imputation model only to estimate the distance between donors and donees, imputations are less sensitive to misspecification of the regression model. As Li and Duan (1989) have shown, under appropriate conditions, when the link function is misspecified in a linear regression, estimates of regression coefficients are consistent up to a multiplicative scalar. Therefore, even under link misspecification (e.g., assuming an identity link when the variable to be imputed is binary) we can still consistently estimate distances between donors and donees. This property has lead Schenker and Taylor (1996) to note that predictive mean matching methods have a built in robustness to misspecfication of the link function.

We have described a SAS macro **MIDAS** for imputing missing data using a predictive mean matching hot-deck. **MIDAS** is a very flexible imputation procedure that can handle data in a variety of formats. Because the imputation models are implicit, it is not necessary to specify a parametric model for each variable to be imputed. Variables are imputed one-at-a-time which allows for each imputation model to condition on a different set of variables. In addition, **MIDAS** allows the user to investigate the impact of different assumptions regarding the missing data mechanism on post-imputation inferences.

Acknowledgments

The authors wish to thank Nicholas Horton for providing SAS code to create the analytic HCUP data set and two anonymous reviewers whose comments greatly improved this manuscript. Siddique's research was supported by grant F31 MH066431-01A1 from the National Institute of Mental Health. Table 3 is reprinted with permission from The American Statistician (Copyright 2007 by the American Statistical Association. All rights reserved).

References

Abayomi K, Gelman A, Levy M (2008). "Diagnostics for Multivariate Imputations." *Applied Statistics*, **57**, 273–291.

Belin TR, Diffendal GJ, Mack S, Rubin DB, Schafer JL, Zaslavsky AM (1993). "Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation." *Journal of the American Statistical Assocation*, **88**, 1149–1166.

Collins LM, Schafer JL, Kam CM (2001). "A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures." *Psychological Methods*, **6**, 330–351.

Cytel Inc (2006). ***LogXact** 8: Discrete Regression Software Featuring Exact Methods.* Cytel Software Corporation, Cambridge, MA. URL http://www.cytel.com/.

Daniels MJ, Hogan JW (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis.* Chapman & Hall/CRC, New York.

Demirtas H, Arguelles LM, Chung H, Hedeker D (2007). "On the Performance of Bias-Reduction Techniques for Variance Estimation in Approximate Bayesian Bootstrap Imputation." *Computational Statistics and Data Analysis*, **51**, 4064–4068.

Demirtas H, Schafer JL (2003). "On the Performance of Random-Coefficient Pattern-Mixture Models for Non-Ignorable Drop-Out." *Statistics in Medicine*, **22**, 2553–2575.

Harrell Jr FE (2008). ***Hmisc**: Harrell Miscellaneous.* R package version 3.5-2, URL http://CRAN.R-project.org/package=Hmisc.

Healthcare Cost and Utilization Project – HCUP (2003). *Overview of the HCUP Kids' Inpatient Database (KID), 2000.* Agency for Healthcare Research and Quality, Rockville, MD. URL http://www.ahrq.gov/data/hcup/.

Honaker J, King G, Blackwell M (2008). ***Amelia**: Amelia II – A Program for Missing Data.* R package version 1.1-33, URL http://CRAN.R-project.org/package=Amelia.

Horton NJ, Kleinman KP (2007). "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models." *The American Statistician*, **61**, 79–90.

Insightful Corp (2003). *S-PLUS Version 6.2*. Seattle, WA. URL `http://www.insightful.com/`.

Li KC, Duan N (1989). "Regression Analysis Under Link Violation." *The Annals of Statistics*, **17**, 1009–1052.

Little RJ (1988). "Missing-Data Adjustments in Large Surveys." *Journal of Business & Economic Statistics*, **6**, 287–301.

Little RJ, Rubin DB (2002). *Statistical Analysis with Missing Data*. 2nd edition. John Wiley & Sons, Hoboken.

Little RJ, Schluchter MD (1985). "Maximum Likelihood Estimation for Mixed Continuous and Categorical Data with Missing Values." *Biometrika*, **72**, 497–512.

Liu C, Rubin DB (1998). "Ellipsoidally Symmetric Extensions of the General Location Model for Mixed Categorical and Continuous Data." *Biometrika*, **85**, 673–688.

Meng XL (1994). "Multiple-Imputation Inferences with Uncongenial Sources of Input." *Statistical Science*, **9**, 538–573.

Raghunathan TE, Lepkowski JM, Hoewyk JV, Solenberger P (2001). "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." *Survey Methodology*, **27**, 85–95.

Raghunathan TE, Solenberger PW, Hoewyk JV (2002). ***IVEware**: Imputation and Variance Estimation Software User Guide*. Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan. URL `http://www.isr.umich.edu/src/smp/ive/`.

Royston P (2005). "Multiple Imputation of Missing Values: Update of **ice**." *The Stata Journal*, **5**(4), 527–536.

Rubin DB (1976). "Inference and Missing Data." *Biometrika*, **63**, 581–592.

Rubin DB (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Hoboken.

Rubin DB, Schenker N (1986). "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association*, **81**, 366–374.

Rubin DB, Schenker N (1991). "Multiple Imputation in Health-Care Databases: An Overview and some Applications." *Statistics in Medicine*, **10**, 585–598.

Rubin DB, Stern HS, Vehovar V (1995). "Handling 'Don't Know' Survey Responses: The Case of the Slovenian Plebiscite." *Journal of the American Statistical Assocation*, **90**, 822–828.

SAS Institute Inc (2003). *The SAS System, Version 9.1.* Cary, NC. URL http://www.sas.com/.

Schafer JL (1997). *Analysis of Incomplete Multivariate Data.* Chapman & Hall/CRC, New York.

Schafer JL, Graham JW (2002). "Missing Data: Our View of the State of the Art." *Psychological Methods*, **7**, 147–177.

Schenker N, Taylor JM (1996). "Partially Parametric Techniques for Multiple Imputation." *Computational Statistics and Data Analysis*, **22**, 425–446.

Siddique J, Belin TR (2008a). "Multiple Imputation Using an Iterative Hot-Deck with Distance-Based Donor Selection." *Statistics in Medicine*, **27**, 83–102.

Siddique J, Belin TR (2008b). "Using an Approximate Bayesian Bootstrap to Multiply Impute Nonignorable Missing Data." *Computational Statistics & Data Analysis*, **53**, 405–415.

van Buuren S, Oudshoorn C (2007). ***mice**: Multivariate Imputation by Chained Equations.* R package version 1.16, URL http://CRAN.R-project.org/package=mice.

Wachter KW (1993). "Comment on Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation." *Journal of the American Statistical Assocation*, **88**, 1161–1163.

Affiliation:

Juned Siddique
Department of Preventive Medicine
Northwestern University
680 North Lake Shore Drive, Suite 1102
Chicago, IL 60611, United States of America
E-mail: siddique@northwestern.edu
URL: http://www.preventivemedicine.northwestern.edu/siddique.htm

Ofer Harel
Department of Statistics
University of Connecticut
215 Glenbrook Road Unit 4120
Storrs, CT 06269, United States of America
E-mail: oharel@stat.uconn.edu
URL: http://www.stat.uconn.edu/~oharel/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 29, Issue 9 *Submitted:* 2008-09-10
February 2009 *Accepted:* 2009-01-28