# A Flexible Method to Apply Multiple Imputation Using SAS/IML® Studio

Lisa M. Lix[1], Xue Yao[1]
[1]Department of Community Health Sciences, University of Manitoba, Winnipeg, MB, Canada

## Abstract

Multiple imputation methods are widely used for missing data problems in various scientific fields. Imputation methods can also be applied to measurement error problems, which arise frequently in many data-analytic problems. SAS/STAT® software offers the MI and MIANALYZE procedures for creating and analyzing multiple imputation data. PROC MI can be used to impute continuous or categorical variables with a monotone missingness pattern and continuous variables with an arbitrary missingness pattern. This paper provides a flexible imputation method developed using SAS/IML® Studio for categorical variable with an arbitrary missingness pattern. This method expands the SAS analyst's ability to apply multiple imputation methods to a wide variety of variables.

## Introduction

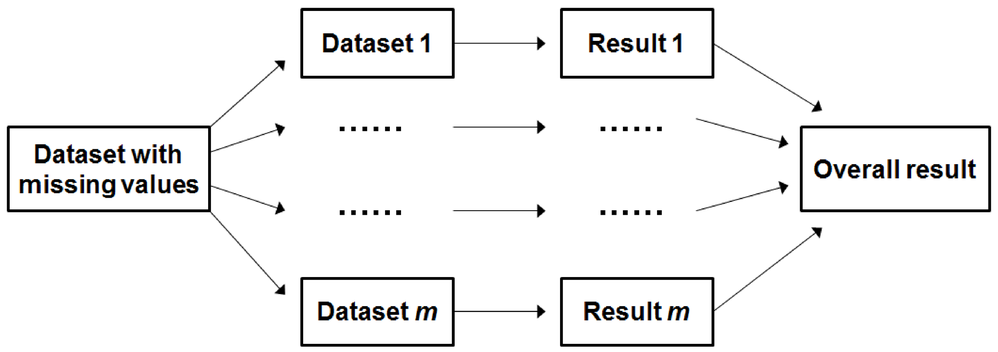


**Figure 1. Schematic Diagram of the Multiple Imputation Model**

| Pattern of Missingness | Type of Imputed Variables | Recommended Methods |
|---|---|---|
| Monotone | Continuous | Regression |
| | | Predicted mean modeling |
| | | Propensity score |
| Monotone | Classification (Ordinal) | Logistic regression |
| Monotone | Classification (Nominal) | Discriminant function method |
| Arbitrary | Continuous | MCMC full-data imputation |
| | | MCMC monotone-data imputation |

**Table 1. Imputation Methods Available in PROC MI**

## Purpose

To develop a **flexible multiple imputation method** for categorical variables (i.e., binary, nominal, ordinal) with an arbitrary missingness pattern.

This method is developed using SAS/IML® Studio.

## Flexible Multiple Imputation Method

### Step 1: Prepare the Data

Prepare the data to model the probability of missing observations in a validation dataset, which contains the true value of the variable and multiple potential predictors of the true value.

In SAS/IML® Studio, the data can be created from:

- SAS/IML matrices
- SAS server data sets
- SAS/IML® Studio data
- External files (e.g., Microsoft Excel)

### Step 2: Use SAS/STAT to Build the Imputation Model

Choose an imputation model based on data exploration and the expected relationship between observed and missing values in the validation data set.

Examples of SAS procedures to develop the imputation model:

- GENMOD (generalized linear model)
- MCMC (Bayesian model)
- CATMOD (linear models for functions of categorical data)
- PROBIT (models with probit, logit or complementary log-log link functions)
- PRINQUAL (principal components analysis)

Save the parameter estimates and their variances and covariances using OUTEST or OUTPOST statements.

### Step 3: Generate Multiple Parameters

Assume the parameters asymptotically follow a multivariate normal distribution.

We developed the betagen module in SAS/IML to generate multiple values of the parameter estimates.

### Step 4: Create Multiple Complete Datasets

Impute the missing variable by replacing the parameters of the imputation model with each vector of coefficients.

The imputed plausible values for the missing variable are saved into a matrix by the number of imputations (e.g., 10).

### Step 5: Analyze Multiple Complete Datasets

The complete data can be analyzed by virtually any technique that would be appropriate to obtain the parameter estimates of interest, such as:

Means:

- UNIVARIATE (DATA=dataset)
- CORR (DATA=COV dataset)

- Regression Coefficients
  - REG (DATA=EST dataset)
  - MIXED (PARMS= and COVB=datasets)
  - GENMOD (ParameterEstimates= and CovB=datasets)
  - GLM (ParameterEstimates= and InvXPX=datasets)
- Correlation Coefficients
  - CORR (FisherPearsonCorr= dataset)

### Step 6: Use PROC MIANALYZE to Combine Results

The MIANALYZE procedure reads parameter estimates and associated standard errors or covariance matrices for each complete dataset.

The combined results from different imputed datasets are used to conduct valid statistical inferences that reflect the uncertainty in the data due to missing values.

## EXAMPLE

| VALIDATION INDICATOR | TRUE DISEASE (*Y*) | OBSERVED DISEASE (*U*) | *X*1 | *M*1 | *X*2 | *M*2 |
|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0.32 | 0.21 | 1 | 1 |
| 1 | 0 | 0 | 2.22 | 1.92 | 1 | 1 |
| 0 | ? | 0 | ? | -1.95 | ? | 0 |
| 1 | 0 | 0 | 0.99 | 2.89 | 1 | 1 |
| 1 | 0 | 0 | 1.68 | 0.09 | 1 | 1 |
| 1 | 1 | 0 | -0.85 | -2.35 | 1 | 1 |
| 0 | ? | 0 | ? | 1.23 | ? | 0 |
| 1 | 0 | 0 | 1.82 | 1.17 | 0 | 1 |
| 0 | ? | 0 | ? | 2.70 | ? | 0 |
| 1 | 1 | 0 | 0.22 | 0.05 | 0 | 1 |
| 1 | 0 | 0 | 2.49 | 2.03 | 0 | 0 |
| 0 | ? | 0 | ? | -0.63 | ? | 1 |
| 1 | 0 | 0 | 0.36 | 0.50 | 0 | 0 |
| … | … | … | … | … | … | … |

**Figure 2. Illustration of a health dataset with measurement error in selected variables**

- True disease status *Y*, the variable of interest, is observed in a validation dataset, the missing *Y* values need to be imputed in the rest of the data set.
- The model for true disease status is developed in the validation dataset using one or more predictor variables (e.g., *M*1 and *M*2).
- The predicted probability of having the disease is used to conduct the multiple imputation.

## CONCLUSIONS

- SAS/IML® Studio provides an integrated development environment for multiple imputation methods by combining:
  - the flexibility of SAS/IML,
  - analytical power of SAS/STAT procedures, and
  - data manipulation capabilities of DataObject class in SAS/IML® Studio.
- The **betagen module** was developed to generate multiple values based on an imputation model generated from observed data.
- Breaking the multiple imputation process into a series of linked steps improves the flexibility of the process for a variety of models.

## ACKNOWLEDGEMENTS

**This research was supported by funding from the Canadian Institutes of Health Research and the Saskatchewan Health Research Foundation.**

**Contact Authors:**
**Xue Yao** Xue.Yao@med.umanitoba.ca
**Lisa M. Lix** Lisa.Lix@med.umanitoba.ca

SAS® GLOBAL FORUM