SAS/STAT Software Enhancements

Multiple Imputation for Missing Data

Overview

SAS/STAT software, Version 8, introduces the experimental MI and MIANALYZE procedures for creating and analyzing multiply imputed data sets for incomplete multivariate data. Multiple imputation provides a useful strategy for dealing with data sets with missing values. Instead of filling in a single value for each missing value, Rubin's (1987) multiple imputation procedure replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. These multiply imputed data sets are then analyzed by using standard procedures for complete data and combining the results from these analyses. No matter which complete-data analysis is used, the process of combining results from different imputed data sets is essentially the same. This results in statistically valid inferences that properly reflect the uncertainty due to missing values.

The MI procedure is a multiple imputation procedure that creates multiply imputed data sets for incomplete *p*-dimensional multivariate data. It uses methods that incorporate appropriate variability across *m* imputations. Once the *m* complete data sets are analyzed using standard SAS/STAT procedures, PROC MIANALYZE can be used to generate valid statistical inferences about these parameters by combining the results.

Introduction

Most SAS statistical procedures exclude observations with any missing variable values from an analysis. These observations are called incomplete cases. While using only complete cases has its simplicity, you lose information in the incomplete cases. This approach also ignores the possible systematic difference between the complete cases and incomplete cases, and the resulting inference may not be applicable to the population of all cases, especially with a smaller number of complete cases.

Some SAS procedures use all the available cases in an analysis, that is, cases with available information. For example, PROC CORR estimates a variable mean by using all cases with nonmissing values on this variable, ignoring the possible missing values in other variables. PROC CORR also estimates a correlation by using all cases with nonmissing values for this pair of variables. This may make better use of the available data, but the resulting correlation matrix may not be positive definite.

Another strategy is single imputation, in which you substitute a value for each missing value. Standard statistical procedures for complete data analysis can then be used with the filled-in data set. For example, each missing value can be imputed from the variable mean of the complete cases, or it can be imputed from the mean conditional on observed values of other variables. This approach treats missing values as if they were known in the complete-data analyses. Single imputation does not reflect the uncertainty about the predictions of the unknown missing values, and the resulting estimated variances of the parameter estimates will be biased towards zero.

Instead of filling in a single value for each missing value, a multiple imputation procedure (Rubin 1987) replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. The multiply imputed data sets are then analyzed by using standard procedures for complete data and combining the results from these analyses. No matter which complete-data analysis is used, the process of combining the results from different data sets is essentially the same.

SAS/STAT procedures implements multiple imputation inferences in three distinct phases:

- Create *m* multiply imputed complete data sets using the MI procedure
- Analyze the *m* complete data sets by using standard procedures such as PROC REG or PROC GLM.
- Generate valid statistical inferences about the parameters of interest by combining the results using the MIANALYZE procedure.

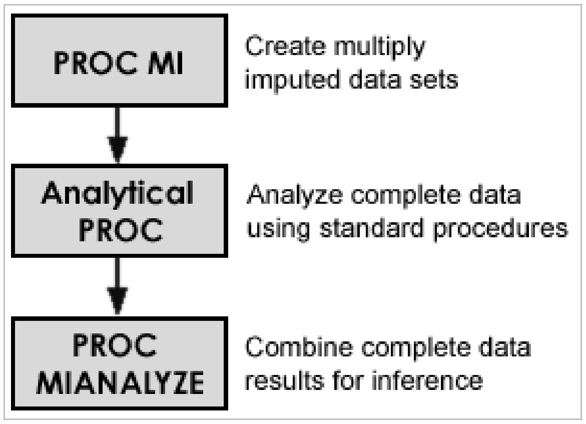
PROC MI
Create multiply imputed data sets
Analytical PROC
Analyze complete data using standard procedures
PROC MIANALYZE
Combine complete data results for inference

Figure 1. The Multiple Imputation Process using SAS Software

Imputation Mechanisms

The SAS multiple imputation procedures assume that the missing data are missing at random (MAR), that is, the probability that an observation is missing may depend on the observed values but not the missing values. These procedures also assume that the parameters θ of the data model and the parameters ϕ of the missing data indicators are distinct. That is, knowing the values of θ does not provide any additional information about ϕ, and vice versa. If both MAR and the distinctness assumptions are satisfied, the missing data mechanism is said to be ignorable.

The MI procedure provides three methods for imputing missing values and the method of choice depends on the type of missing data pattern. For monotone missing data patterns, either a parametric regression method that assumes multivariate normality or a nonparametric method that uses propensity scores is appropriate. For an arbitrary missing data pattern, a Markov chain Monte Carlo (MCMC) method that assumes multivariate normality can be used.

Regression Method

In the regression method, a regression model is fitted for each variable with missing values, with the previous variables as covariates. Based on the resulting model, a new regression model is then simulated and is used to impute the missing values for each variable.

Propensity Score Method

The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates. In the propensity score method, a propensity score is generated for each variable with missing values to indicate the probability of the observation being missing. The observations are then grouped based on these propensity scores, and an approximate Bayesian bootstrap imputation is applied to each group.

MCMC Method

In MCMC, one constructs a Markov chain long enough for the distribution of the elements to stabilize to a common, stationary distribution. By repeatedly simulating steps of the chain, it simulates draws from the distribution of interest.

In Bayesian inference, information about unknown parameters is expressed in the form of a posterior distribution. MCMC has been applied as a method for exploring posterior distributions in Bayesian inference. That is, through MCMC, one can simulate the entire joint distribution of the unknown quantities and obtain simulation-based estimates of posterior parameters that are of interest.

Assuming that the data are from a multivariate normal distribution, data augmentation is applied to Bayesian inference with missing data by repeating a series of imputation and posterior steps. These two steps are iterated long enough for the results to be reliable for a multiply imputed data set (Schafer 1997). The goal is to have the iterates converge to their stationary distribution and then to simulate an approximately independent draw of the missing values.

Release 8.2

Release 8.2 of SAS/STAT software includes the second experimental releases of the MI and MIANALYZE procedures. Additions to PROC MI include the TRANSFORM statement to transform variables before performing the imputation, autocorrelation and iteration plots, a monotone-data MCMC method to impute just enough values to achieve a monotone missing pattern for the imputed data, and the EM statement to derive the MLE and related EM results.

For More Information

For more information about the new multiple imputation procedures and other analytical software in the SAS System, visit the Statistics and Operations Community website at **www.sas.com/statistics/.** The paper "Multiple Imputation for Missing Data: Concepts and New Development" is also available from this site.

Reference

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons, Inc.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, New York: Chapman and Hall