ST-14
Handling Missing Data with Multiple Imputation Using PROC MI in SAS

Hossein N. Yarandi, PhD, University of Florida, Gainesville, Florida

ABSTRACT

The multiple imputation was developed as a general method for inference with missing data. Instead replacing the missing observation with a single value, multiple imputation method replaces each missing value with multiple plausible values. PROC MI in SAS creates multiply imputed data sets for incomplete multivariate data. This study reviews multiple imputation as an analytic strategy for missing data and applies PROC MI to impute missing data in a Medical Expenditure Panel Survey.

INTRODUCTION

In the data analysis phase of research, missing values present a challenge to investigators. In most applied research studies, incomplete data is the rule and not the exception. In particular, for the analysis of survey data, accommodation of the incomplete data is critical to making valid inferences. Almost invariably, the data available to social scientists display one or more characteristics of missing information. Response rates in surveys, for instance, have been found to range between 13 to 95 percent (Little and Rubin, 1992). Even though reasons for no response are varied, most frequently they reflect the unwillingness of respondents to provide information on undesirable social behaviors and on issues that they consider as private. Besides these, poor research designs often lead to ambiguous and poorly structured survey questions, which create a recipe for low response (Fay, 1986; Rubin, 1985). In addition, longitudinal surveys suffer from incompleteness due to attrition resulting from death and relocation (Little, 1985).

The problems of analyzing data with missing values have been reviewed extensively in the scientific literature (Little, 1992; Little & Rubin, 1989; Rubin, 1987; Shafer, 1997). Common approaches for addressing missing data generally include pairwise and casewise deletion, mean imputation, nearest neighbor imputation, deductive imputation, hot deck imputation, regression imputation, random (or stochastic) regression imputation, and propensity matching (Rubin, 1986; Little and Rubin, 1987). Through these methods, also known as single imputation, wherein a single value is imputed for each missing observation. Single imputation methods have been the subject of increasing criticism with respect to their tendency to underestimate standard errors, overstate statistical significance, and introduce bias (Rubin and Schenker, 1986). However, if the proportion of missing values is small (less then 5%), then a simple imputation method may be considered to be accurate.

An alternative method for addressing incomplete data is multiple imputation (MI). Through this method, we can replace 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. MI does not attempt to estimate each missing value through simulated values but rather to represent a random sample of the missing values. This process results in valid statistical inferences that more accurately reflect the uncertainty due to missing values.

TYPES OF MISSING DATA

Missing values in any research occurs for a variety of reasons. Rubin (1976) introduces the term “missing completely at random” (MCAR) to describe missing data when the probability that a missing response is independent of the true, but unobserved values for that variable. In addition, MCAR also assumes the probability that the missing response is independent of the observed covariates in the data set. That is, data are missing totally accidentally. An example of a MCAR process is one in which respondents decide whether to answer survey questions on the basis of coin flips. A more realistic assumption is when data are “missing at random” (MAR), in which case the data are dependent on the observed covariates, but

the probability that a missing response is independent of the unobserved values for that variable (Little and Rubin, 1989). For example, if the subjects with higher income are more likely to refuse to answer an income question, then the process will be a MAR as long as the respondents complete other variables such as sex, age, race, and education. MAR is sometimes called ignorable non-response, meaning that a statistical model can explain the non-response, hence, the non-response can be ignored after a model accounts for it. MAR assumptions can be made to fit the data by including more variables in the imputation process to predict the pattern of missingness.

Another form of missingness occurs when the probability of missing observations depend on the unobserved values of the missing responses. This process is known as non-ignorable. That is, values that are missing are systematically different than those observed even among subjects with similar covariates. In this situation, the missingness is said to be not missing at random (NMAR). For instance, when high-income respondents are more likely to refuse to answer survey questions about income and when other variables in the data set cannot predict which respondents have high income.

The performance of different methods of analyzing incomplete data under MCAR, MAR, or NMAR depends upon the ultimate goals of the analysis. However, by definition, the presence or absence of NMAR can never be demonstrated using only the observed data. If we assume that NMAR exits, the missing data cannot be predicted without bias from observed covariates and no general method for correction is available.

MUTIPLE IMPUTATION

Multiple imputation (MI) involves imputing m values for each missing item and creating m ($m > 1$) complete data sets. Across these complete data sets, the observed values are the same, but the missing values are filled in with different imputations to reflect uncertainty levels (Little and Rubin, 1989; Landerman, Land and Pieper, 1997). The reasoning behind the m replications of imputed values is to create m complete data matrices each of which is to be analyzed by standard complete data methods. This approach retains the advantage of single imputation by using a complete data set and thus allowing standard statistical methods to be utilized. At the same, however, by allowing more than one value on a missing variable to be estimated, MI corrects for sampling variability and thus improves upon single imputation techniques, which uses only a single value. The imputed m values on the variable of interest can therefore be aggregated to produce inferential results. In addition, random error in the imputation process yields approximately unbiased estimates of all parameters, which no deterministic method can perform. Also, repeated imputation allows for good estimates of the standard errors.

In using the MI, the question arises as to how much replications (m) ought to be applied in order achieve biased estimates. Comparing simulation models based on single and multiple imputation techniques, Little and Rubin (1987) and Rubin and Schenker (1986) have demonstrated that even in extreme cases where the proportion of missing information constitute about a third of the data set, no more than 5 replicates ($m \leq 5$) of the model provides efficient estimates. In practice, the multiple imputation process involves: (1) imputing missing values using a random variation model, (2) producing usually between 3-5 complete data sets, (3) performing the desired analysis on each compete data set, (4) averaging the values of the parameter estimates to produce a single point estimate, and (5) calculating the standard errors.

Implementing MI requires that the data are MAR, conditional on the imputation model. That is, it is assumed that missing data values carry no information about probabilities of missingness. This assumption is mathematically convenient because it allows one to eschew an explicit probability model for non-response. In some applications, however, ignorability may seem artificial or implausible. With attrition that is likely to occur in a longitudinal study, for example, it is possible that subjects drop out for reasons related to current data values. However, the MAR assumption can be made more realistic by including additional informative variables. Even when the MAR condition is satisfied, creating random imputations that result in unbiased estimates of the desired parameters is not always easy or straightforward. In addition, MI assumes that the variables are jointly multivariate

normal. This model obviously is an approximation, as few data sets have variables that are all continuous and unbounded, much less multivariate normal. Yet researchers have found it to work as well as more complicated alternatives specially designed for categorical or mixed data sets (Schafer, 1997; Rubin and Schenker, 1991). Transformations and other procedures can be employed to improve the fit of the model. However, inferences based on MI can be robust to departure from the multivariate normality if the amount of missing information is not large (Schafer, 1997).

THE MI PROCEDURE

The new MI procedure (PROC MI) creates multiply imputed data sets for incomplete multivariate data. It uses methods that incorporate appropriate variability across the *m* imputations. The method of choice depends on the patterns of missingness. Once the *m* complete data sets are analyzed using standard SAS procedures, the new MIANALYZE procedure can be employed to generate valid statistical inferences about these parameters by combining results from the *m* analyses. These two procedures are available in experimental form in Release 8.2 of the SAS System (http://www.sas.com/rnd/app/da/new/802ce/stat/).

Three methods are available in the MI procedure using SAS. (1) Markov Chain Monte Carlo (MCMC) is a collection of methods for simulating random draws from nonstandard distributions via Markov chains (Schafer, 1997). MCMC is used to create a small number of independent draws of the missing data from a predictive distribution, and these draws are then used for multiple-imputation inference. With an arbitrary missing data pattern, we can often use the MCMC method, which creates multiple imputations by drawing simulations from a Bayesian predictive distribution for normal data. Another use the MCMC is to impute enough values to make the missing data pattern monotone. Then, we can use a more flexible imputation method. (2) A regression model is fitted for each variable with missing values, with the previous variables as covariates. Based on the fitted regression coefficients, a new regression model is simulated from the posterior predictive distribution of the parameters and is used to impute the missing values for each variable (Rubin 1987). The process is repeated sequentially for variables with missing values. (3) A propensity score method, which is generally defined as the conditional probability of assignment to a particular treatment given a vector of observed covariates, can be used (Rosenbaum and Rubin 1983). In the propensity score method, for each variable with missing values, a propensity score is generated for each observation to estimate the probability that the observation is missing. The observations are then grouped based on these propensity scores, and an approximate Bayesian bootstrap imputation is applied to each group.

AN EXAMPLE

The data set for this example are obtained from the Medical Expenditure Panel Survey (MEPS) that is cosponsored by the Agency for Healthcare Research and Quality (AHRQ), and the National Center for Health Statistics (NCHS). MEPS provides nationally representative estimates of health care use, its expenditure, sources of payment, and insurance coverage for non-institutionalized population within the United States. MEPS also includes a national representative survey of nursing homes and their residents, and is comprised of four component surveys: the household component, the medical provider component, the insurance component, and the nursing home component. In addition, it allows for health services research intended to inform health care policy.

Designed to capture the changing dynamics of health care delivery and its insurance systems, MEPS began in 1996. An important feature of MEPS is that it utilizes a panel survey design in which minority and low-income households are oversampled. Medical expenditure data are collected at both the individual and household levels through a prescreening telephone interview, a mailed questionnaire, and a telephone follow-up interview for non-respondents. The data reflect demographic characteristics, income, employment, health conditions, health status, access and use of health care services, charges and payments, satisfaction with care, and health insurance coverage.

From the MEPS survey a sample of n = 759 single mothers was selected for this study. The variables considered were difficulty to obtain health insurance (1 = Yes, 0 = No), race (0 = Nonwhite, 1 = White), employment (1 = Yes, 0 = No), number of children,

years of formal education, and annual income level which was categorized as negative or poor, near poor, low income, middle income, and high income. Four dummy variables were generated to represent the five income categories. None of the variables had missing values. First, a logistic regression was performed using the variable, difficulty to obtain health insurance, as the dichotomous dependent variable and the remaining variables were employed as the independent variables. Second, missing values were induced for middle and high-income subjects using MAR definition. About 30% of the middle income and 33% of the higher income respondents were marked as missing. Third, a logistic regression was performed excluding all cases on which data were missing (casewise deletion). Forth, MI was implemented on the entire data set by using the PROC MI. Five imputed data sets were generated and logistic regression was performed on each of these imputed data sets. The coefficients shown in Table 1 are the means of the five estimates using PROC MIANALYZE.

As indicated in Table 1, the estimated coefficients and the corresponding standard errors using the complete data set and the imputed date set revealed similar values. That is, MI procedure produced reasonably good results. On the other hand, there were larger discrepancies in the estimated coefficients and standard errors between the original data sets and casewise deletion.
In addition, casewise deletion resulted in opposite signs of the estimated coefficients for middle- and high-income respondents compare to the original data set. The standard errors using the casewise deletion were greater than the other two methods.

Researchers should consider MI methods to help alleviate problems caused by survey non-response and missing data. A major advantage in the use of MI methods is that the variation between the m imputations reflects the uncertainty with which the missing values can be estimated from the observed values. The previous example showed that the MI method could be successfully implemented, under certain conditions, for applied research projects using PROC MI.

Table 1. Estimates for Coefficients in Logistic Regression

	Original Data Set		Casewise Deletion		Imputed Data Set	
Parameter	Estimate	Std Error	Estimate	Std Error	Estimate	Std Error
Intercept	3.0465	0.4124	3.6322	0.6913	3.0174	0.4176
Employment	0.3538	0.2318	0.8744	0.3166	0.3610	0.2327
Children	-0.1372	0.0839	-0.2218	0.1568	-0.1398	0.0848
Race	-0.7104	0.2234	-1.1223	0.3260	-0.7154	0.2235
Income						
Near Poor	-0.6436	0.3256	-1.0479	0.4358	-0.6493	0.3270
Low	-0.5860	0.2830	-0.9070	0.3736	-0.5918	0.2836
Middle	0.0704	0.3283	-0.3864	0.4558	0.0512	0.3552
High	0.2713	0.4770	-0.2506	0.6747	0.3981	0.7226
Education	-0.0821	0.0314	-0.0940	0.0680	-0.0768	0.0328

REFERENCES

Fay, R. (1986). Causal models for patterns of non-response. Journal of the Statistical Association of America, 91(394), 355-365

Landerman, L., Land, K., and Pieper, C. (1997) An empirical evaluation of the predictive mean matching method for imputing missing values. Sociological Methods & Research 26, 3-33.

Little, R. (1992). Regression with missing X's: a review. Journal of the American Statistical Association, 87, 1227-37.

Little, R. (1985). Modelling drop-out mechanism in repeated measures studies. Journal of the American Statistical Studies, 90(43), 1112-1121.

Little, R., & Rubin, D. (1989). Statistical analysis with missing data. New York: Wiley.

Rosenbaum, P. and Rubin, D. (1983), The central role of the propensity score in observational studies for causal effects, Biometrika, 70, 41-55.

Rubin, D. (1976). Inference and missing data, Biometrika, 63, 581-592.

Rubin, D. (1987). Multiple imputation for non-response in surveys. New York: Wiley.

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

Rubin, D. & Schenker, N. (1991). Multiple imputation in health-care databases: An overview and some applications. Statistics in Medicine, 10, 585-598.

Rubin, D. & Schenker, N. (1986). Multiple imputation from random samples with ignorable non-response. Journal of the American Statistical Association, 81(394), 366-374.

Shafer, J. (1997). Analysis of incomplete multivariate data. London: Chapman and Hall

CONTACT INFORMATION

Hossein N. Yarandi, PhD
Associate Professor
University of Florida
Campus Box 100187
Gainesville, FL 32610-0187

Voice: (352) 846 - 0658
Fax: (352) 846 - 1624
E-mail: yarandi@ufl.edu