A57

Broad Scale Missing-Value Imputation With Iterative Binary Partitioning

Dan Steinberg
San Diego State University

Richard Carson
University of California, San Diego

Leo Breiman
University of California, Berkeley

Introduction

Most large databases contain numerous missing data points. This is particularly true in the social sciences where missing values are common in the two major forms of data collection, surveys and administrative records. Missing values occur in questionnaires when respondents fail to answer some questions or when screening procedures set implausible values for observations as missing. Missing values occur in administrative records when information is not available at the time of original data entry or, as frequently happens, is never recorded. Frequently, a substantial fraction of the observations in a database have one or more missing values. Thus, a good method for handling observations with missing values is needed to effectively summarize the information in a large database.

The most common empirical practice for dealing with missing data is to drop any observation with a missing value on any variable of interest. This common practice, however, results in inconsistent estimates of both univariate statistics and the relationships between the variables except in the fairly unusual case when variable values are *missing completely at random* (Little and Rubin, 1987). Even in this case, where missing values are said to be *ignorable*, dropping observations generally results in inefficient estimates, particularly if the fraction of missing data is large. For this reason, ad hoc methods are often used in empirical practice, such as setting a missing value on a variable equal to the variable's mean or to the last valid value from a hot deck.[1] Under typical circumstances, however, variable values are not missing at random. Thus, it is useful to identify two different situations that differ on whether a variable's values are missing at random or not at random conditional on a set of observed covariates.

The first case is amenable to imputing consistent estimates for the missing values using a variety of classical and Bayesian techniques (Little and Rubin, 1987; Gelman *et al.*, 1995). To impute missing values in this case, one specifies a statistical model, typically in a regression or maximum likelihood framework, to predict the missing values. The most popular classical and Bayesian approaches are based upon the well-known EM algorithm where some missing values (*e.g.*, those on the first variable of interest) are predicted first on the basis of complete observations and then, conditional on those imputed values and the observed data, other missing values are imputed. The process is continued with the original imputed values being progressively replaced with more refined estimates until some convergence criterion is met. This paper proposes an alternative to these methods based upon binary partitioning approach of CART (Breiman, *et al.*, 1984).

Consistent estimates of the missing values can also be obtained in the second situation, but often with considerably more difficulty because the mechanism causing the missing values is tied to the unobserved error distribution and often involves unknown censoring or truncation processes. This second situation is often referred to in econometrics as the sample selection bias issue (Heckman, 1979). We do not explicitly deal with this case here except to note that using an additional covariate can change a (b) situation to an (a) situation, as can knowing the functional relationship (up to an estimable parameter vector) between the variables of interest. As a consequence, we believe that more situations will fall into (a) rather than into (b) when using a CART-based missing value imputation approach than when using an EM-based approach due to CART's ability to effectively search over large sets of possible covariates and to uncover highly non-linear relationships.

Characteristics of An Ideal Missing-Value Imputation Procedure

An ideal missing valuation imputation procedure should have six characteristics. It should produce estimates for the missing values that are (1) unbiased and (2) reasonably efficient. The procedure should work well with (3) both continuous and categorical variables and mixtures of these two types of variables. Further, the estimates for the missing values should be fairly insensitive to specific modeling assumptions, particularly those involving (4) error distributions and (5) the functional relationships between variables. Finally, the procedure should (6) be easily automated for a large number of variables and not require substantial hand tuning or individual functional form specification for each variable of interest.

Ad hoc imputation measures such as replacing missing values with the variable's mean meet only (4), (5), and (6). In general, estimates from such procedures tend to use available information very inefficiently and can introduce substantial bias into the parameter estimates of interest. Regression and EM-based approaches also usually fail (4), (5), and (6), although robust variants of these approaches that solve some of the issues with (4) are available. Automated versions of the EM approach, such as the procedure contained in the statistical package BMDP, effectively maximize a correlation matrix. This procedure satisfies (6) by giving up (2) and (3) and increasing the sensitivity to (4).

Carson (1984) proposed a missing-value imputation approach based on CART which meets (1), (3), (4), and (5) and also improves on (2) relative to commonly-used ad hoc approaches.[2] CART is widely used as a non-parametric approach to uncovering the relationship between a dependent variable and a large set of possible predictor variables. As a binary partitioning method, it can potentially avoid problems associated with not knowing the structure of the relationship between variables. Like most non-parametric procedures, CART avoids the need to make distributional assumptions but requires a large number of observations to work well. The major drawback of Carson's (1984) proposed use of CART as a method for imputing missing values on a large number of variables was that it required substantially more computational resources than were generally available to researchers at the time. As this limitation is no longer binding even for quite large datasets, we consider the possibility of automating the original CART missing-value approach and improving the efficiency of the estimates of the missing values. We also consider how various imputation options introduced in conjunction with the EM and hot deck missing-value procedures can be incorporated into a CART framework.

A Sketch of Automated CART Missing-Value Imputation

Here we sketch how an automated CART missing value imputation algorithm works. Start with a data matrix $\mathbf{X}$ consisting of $i = 1, ..., n$ rows (observations) and $j = 1, ..., k, ..., m$ columns (variables). $\mathbf{X}$ is allowed to have an arbitrary pattern of missing values, although consistency of the algorithm requires that the missing-at-random conditional-on-available-covariates requirement be met. The algorithm works as follows:

(1) Starting with column vector x_1, estimate a CART tree, dropping all the observations with missing values on x_1 and using the surrogate split feature to handle observations with missing values on the predictor variables x_k ($k \neq 1$). Obtain predicted x_1 for observations with missing x_1 values based upon the terminal nodes defined by the estimation procedure using one of the approaches described in the next section. Repeat this procedure using each column in $\mathbf{X}$ in turn as the dependent variable and the other x_k's as the independent variables.

(2) Now re-estimate a CART tree for each column of $\mathbf{X}$ using as the dependent variable the originally valid x_j values where available or the predicted x_j values from (1) if not. Obtain a new set of predicted x_j values for observations originally missing.

(3) Now re-estimate the CART trees using only originally valid values of x_j and as predictors use the other x_k variables where the values used for these variables are the originally valid x_k values if available or, if originally missing, the predicted x_k from (2). Obtain a new set of predicted values for the x_j originally missing.

(4) Re-estimate CART trees again using originally valid x_j or the predicted x_j from (3) if originally missing. Obtain a new set of predicted x_j.

(5) Repeat (3) and (4) until the convergence criteria are met.

Alternative Terminal Node Imputation Schemes

Conceptually, the terminal nodes from a CART estimation can be thought of as defining an imputation class. Thus, the difference between hot deck

imputation classes and CART imputation classes is that the CART imputation classes are estimated non-parametrically from the data rather than being determined a priori by the researcher on an ad hoc basis. Given the CART imputation classes, it is possible to use a variety of methods to obtain an estimate for a missing value.

The simplest way is to take as the estimate of the missing value a summary statistic such as the mean from the terminal node into which the observation falls. This method improves on the common practice of replacing the missing value of a variable by its unconditional mean by using the conditional mean.

The hot deck exploits user-defined imputation classes and positive spatial or temporal autocorrelation within the imputation class. CART-defined imputation classes should be better on average than those defined by users on the basis of variables thought to be predictive of the variable of interest. Within a CART imputation class, possible positive autocorrelation can be exploited by imputing the value of the observation in the imputation class which is closest according to some measure of distance or time.

The drawback of these approaches, as noted in the EM-based imputation literature (Rubin, 1987), is that they all tend to depress a variable's variability after imputation relative to the true variance of the variable if it had no missing values. This occurs because one is effectively imputing some variant of an expected values. The problem of variance suppression with CART-based missing value imputation can be overcome in similar ways to those proposed for EM-based approaches. Methods include imputing a randomly-chosen valid value from the CART terminal node or using multiple draws from the valid values of the CART terminal node. It is also possible to draw from CART "residuals" or to fit a distribution to the valid values in the imputation class and draw from that distribution.

Discussion

A key issue in implementing the proposed approach is whether it is computationally feasible. To examine this issue we have been using a subset of the U.S. Census Bureau's March 1992 Current Population Survey with 25 variables roughly split between continuous and categorical and 35,000 observations with 20% of observations on each variable randomly set to missing. On a mid-range DEC Alpha with two complete iterations through the data, imputation of all the missing values takes less than 15 minutes. Thus, the proposed procedure seems inside the computationally feasible set for a wide range of problems. The performance of the algorithm in this instance, where most approaches to missing value imputation work well, is quite good, producing unbiased and fairly efficient estimates. We are currently experimenting with how computational requirements change with the number of observations, the number of variables, the number of categorical variables, the number of categories comprising the categorical variables, and the number of times the algorithm is allowed to iterate.

On smaller data sets we have examined the ability of the proposed algorithm to uncover reasonably complex relationships that make the missing-at-random conditional-on-available covariates a reasonable assumption. This is obviously the more common and more important case than the missing-completely-at-random case. In very limited testing, the algorithm appears to work well on these problems, producing estimates that are considerably less biased than most other approaches, including the automated version of the EM algorithm and variations on the hot deck. Much of our future simulation work will attempt to characterize situations where the algorithm is or is not likely to work well.

The iterative structure of this proposed algorithm is similar in many ways to the EM algorithm. Other iterative paths to updating missing variable values than the one proposed here are possible. For instance, it may be desirable in the first step to use predicted values from the CART trees already estimated in predicting subsequent x_j. The relative performance of these various paths is an open research issue. Following the general CART framework, incorporation of priors for categorical data is straightforward and may result in substantial efficiency gains where there is outside information on characteristics of the relevant population. The performance of the proposed procedure with priors should be examined. A number of potential convergence criteria generally can be expressed in terms of changes in either the structure or predictive ability of the CART trees grown for each x_j or in the predicted values of the missing observations. The performance of various crite ria, and more generally, the improvement from additional complete iterations beyond two, is also an open question.

Footnotes

[1] The hot deck is a commonly-used approach for large survey data sets based on user-defined imputation classes and exploiting autocorrelation (typically spatial) within the imputation class. There are many variants of the approach. See Madow *et al.*. (1983) for a discussion.

[2] The approach proposed here is substantively different than one currently embodied in the commercially available version of CART (Steinberg and Colla, 1995), which has built into it a procedure for handling missing values. That procedure does not explicitly impute missing values but rather deals with missing values in a consistent and somewhat more efficient fashion than do most other statistical procedures. CART currently finds the single binary split that maximizes the objective criteria by examining splits based on bivariate pairs of complete cases involving the dependent variable and the allowable set of independent variables. After finding this split, CART determines the other bivariate splits that best mimic the optimal binary split. This information is then used to determine in which direction in the tree to send an observation with a missing value on the independent variable in the optimal split. The list is gone down until there is a "surrogate" split for the observation in which the independent variable does not have a missing value.

References

Breiman, L., J. Friedman, and R. Olshen, and C. Stone (1984), *Classification and Regression Trees* (Pacific Grove, CA: Wadsworth).

Carson, R.T. (1984), "Compensating for Missing Data and Invalid Responses in Contingent Valuation Surveys," in *1984 Proceedings of the Survey Research Section of the American Statistical Association* (Washington: American Statistical Association).

Gelman, , A., J.B. Carlin, H.S. Stern, and D.B. Rubin (1995), *Bayesian Data Analysis* (London: Chapman and Hall).

Heckman, J. (1979), "Sample Selection Bias as Specification Error," *Econometrica*, 47: 153-161.

Little, R.J.A. and D.B. Rubin (1987), *Statistical Analysis with Missing Data* (New York: Wiley).

Madow, W.G., H. Nisselson, I. Olkin, and D.B. Rubin (1983), *Incomplete Data in Sample Surveys* (New York: Academic Press).

Rubin, D.B. (1987), *Multiple Imputation For Non-Response in Surveys* (New York: Wiley).

Steinberg, D. and P. Colla (1995), *CART: A Supplemental Procedure for SYSTAT* (San Diego: Salford Systems).