

Correction of Bias in Imputing Missing Values of Categorical Variables

Ruiguang Song, Kathleen McDavid, Debra L. Hanson, and H. Irene Hall

Centers for Disease Control and Prevention, Atlanta, GA 30333, USA

A commonly used method of multiple imputation for databases with non-monotone patterns of missing data uses the Markov Chain Monte Carlo (MCMC) algorithm, which is based on the multivariate normal model. Because of its wide availability and usefulness with complicated patterns of missing data, the MCMC algorithm has been applied to imputation of categorical variables. A k -level categorical variable is first expressed as a set of $k-1$ dummy-coded binary variables. When the categorical variable is missing, these dummy-coded variables are also missing; the missing data are imputed based on a multivariate normal distribution. The imputed values are rounded to 0 or 1 based on a maximum criterion, e.g. the dummy-coded variable with the highest value is rounded to 1 and the others are rounded to 0. Concerns have been raised about the bias associated with this rounding, but the literature on this issue is limited to binary variables. In this paper, we evaluate the bias resulting from the rounding approach based on the maximum criterion. We propose a method to correct for the rounding bias. Correction factors for imputing binary and three-level categorical variables are provided.