MULTIPLE IMPUTATION FOR MISSING DATA IN MOLECULAR GENETIC STUDIES

Muhanad Akash
Department of Agronomy
Iowa State University
Ames, IA
Gerald O. Myers and Baogong Jiang
Department of Agronomy and Environmental Management
Louisiana State University
Baton Rouge, LA
Barry E. Moser
Department of Experimental Statistics
Louisiana State University
Baton Rouge, LA

Abstract

Until recently, incomplete data were handled primarily either by ignoring subjects with missing information or by substituting plausible values such as means or regression predictions. These approaches may produce answers that are biased, inefficient, or unreliable. Another strategy for handling missing data is multiple imputation (MI) which imputes the missing values multiple times. MI reflects the uncertainty associated with the missing observations, providing unbiased estimates for the parameters of interest and their variances. Our objective in this study is to give a brief overview of missing data handling concepts and several popular methods for handling incomplete data. We then explain how these methods apply to the problem of imputing reasonable values for incomplete quantitative trait loci (QTL) mapping data. Six different methods were applied. This includes propensity score and regression methods for monotone missing patterns of a continuous variable, logistic regression and discriminant function for monotone missing patterns of a binary variable, and markov chain monte carlo (MCMC) full-data imputation and MCMC monotone-data imputation for arbitrary missing patterns of a continuous variable. In this study, propensity score and MCMC full-data imputation gave the best performance with data less than 40% missingness. Both logistic regression and discriminant function methods for binary variables with a monotone pattern of missingness gave correct estimates in most cases. We highly recommend to researchers that they pay attention to the fact that the estimated P-value tends to get higher with an increasing proportion of missingness.