Abstract

Scientific literature lack straight forward answer as to the most suitable method for missing data imputation in terms of simplicity, accuracy and ease of use among the existing methods. Exploration various methods of data imputation is done, and then a robust method of data imputation is proposed. The paper uses simulated data sets generated for various distributions. A regression function on the simulated data sets is used and obtained the residual standard errors for the function obtained. Data are randomly from the set of independent variables to create artificial data-non response and use suitable methods to impute the missing data. The method of Mean, regression, hot and cold decking, multiple, median imputation, list wise deletion, EM algorithm and the nearest neighbour method are considered. This paper investigates the three most common traditional methods of handling missing data to establish the most optimal method. The suitability is hence determined by the method whose imputed data sample characteristic does not vary considerably from the original data set before imputation. The variation is here determined using the regression intercept and the residual standard error. R statistical package has been used widely in most of the regression cases. Microsoft excel is used to determine the correlation of columns in hot decking method; this is because it is readily available as a component of Microsoft package. The results from data analysis section indicated an intercept and R-squared values that closely mirror those of original data sets, suggesting that median imputation is a better data imputation method among the conventional methods. This finding is important from the research point of view, given the many cases of data missingness in scientific research. Finding and using the median is simple and as such most researchers have a ready tool at hand for handling missing data.