A proposal of a method of distinguishing MAR type missing data from MCAR type

Jerzy Korzeniewski
University of Lodz, Lodz, Poland – jurkor@wp.pl

Abstract

Correct assessment of the mechanism responsible for creating missing data is very vital for empirical data sets analysis. If the missing data have random character MCAR (Missing Completely At Random) the analysis is easier because the missing data can be ignored or replaced. The scope of methods dealing with distinguishing the MCAR type from the MAR (Missing At Random) is very limited. Actually, only one test with a rather limited applicability was proposed. In this article we present a proposal of a new method which might be of some help in this task in the case of data sets with a cluster structure. The idea of the method consists in using a measure of the strength of correlation between two variables based on the linear correlation between pairs of objects distances corresponding to the two variables. The measure can be computed in a number of instances e.g. taking into account missing values and treating them as a new value or, omitting missing values. If the strengths of correlation in both instances differ significantly it is a reason for considering the missing values to of the MAR type. The new method is first checked on some binary data sets with a known cluster structure and then on some data sets from the Irvine Data Sets Repository. The results sometimes return very certain answers.

Keywords: class structure, missing data, variable correlation, testing of MCAR type.