Performance Of Imputation Methods Towards Increasing Percentage Of Missing Values

Kenfac Dongmezo Paul Brice
Pan African University
Institute for Basic Sciences, Technology and Innovation (PAUISTI), Kenya
Email – dongmezobrice@gmail.com / brice.dongmezo@students.jkuat.ac.ke

Peter N. Mwita
Machakos University
P.O Box 136-90100; Machakos, Kenya
Email – petermwita@mksu.ac.ke

Kamga Tchwaket Ignace Roger
Institut Sous régional de Statistique et d'Economie Appliquée (ISSEA)
P.O Box 294; Yaoundé, Cameroon
Email – kamignace@yahoo.com

ABSTRACT.
In statistics missing value is an important and very common issue when dealing with collected data. Having a data set without missing values is very rare therefore in case the statisticians have to perform analysis on a dataset, they should think about solution to solve the problem of missing values. Among solution to be undertaken, imputation methods are at the first place. The aim of this paper is to study the performance of five different existing imputation methods (Random Imputation, k Nearest Neighbors Imputation, Mean Imputation, Maximum Likelihood Imputation and Multiple Imputation) from two different perspectives. Firstly, the methods are compared using their ability to estimate the missing observation meaning how close is the data set completed by a given imputation method to the original data set. Secondly, using their ability to estimate some statistics (mean, standard deviation and coefficient of a regression) using the full data set completed by the imputation method. This means that after imputation, important statistics are computed using the completed data set and compared to the original ones if existing. The different comparisons are made using root mean squared error and mean absolute

deviation. Simulation results using simulated data and bootstrap show that Multiple imputation is the best method in completing the data set and in obtaining best estimator of statistics. Random imputation seems better for estimating statistics only while k-Nearest Neighbors for completing data set.
Keywords.Imputation, Errors, Root Mean Squared Error, Mean Absolute error, Performance.