2nd International Conference of Multidisciplinary Approaches (iCMA), 2015
Faculty of Graduate Studies,
University of Sri Jayewardenepura,
Sri Lanka
ISSN: 2386 – 1509
Copyright © iCMA

IDENTIFICATION OF FACTORS AFFECTING MISSING VALUES IN SURVEY DATA AND COMPARISON OF IMPUTATION METHODS

D.A.N. Ranasinghe, N.A.D.N. Napagoda

Department of Mathematical Sciences, Wayamba University of Sri Lanka

arunirans1989@gmail.com

Data of most surveys comprise missing values. Missing values create serious problems for all the parties involved with surveys. Incomplete data occurs when there are no answers for a particular question in the survey questionnaire and it makes analysis of data more challenging. Furthermore, missing values can lead to incorrect decisions. Missing value imputation is used to alleviate these by replacing incomplete data using suitable values.

Imputation methods create a way to obtain accurate results of a survey. This study attempts to investigate the factors affecting missing values and identifies the most appropriate imputation method out of several methods. It helps to increase the accuracy of the survey results and it may help to collect complete data for a survey with the awareness of affecting factors.

Survey data of Annual Surveys of Industries (ASI) collected by the Department of Census and Statistics (DCS) were used for the study. Diffusion of missing values among each identified variables were comprehended and the factors affecting missing values were employed using binary logistic regression model. Moreover, five imputation methods were selected for the comparison such as cold-deck imputation, multiple imputation, expectation maximization, linear interpolation and linear trend at point.The differences between actual and imputed values of each method were calculated and one-way analysis of variance (ANOVA) was used to identify the pairs that have unequal differences.

The results suggest that there are four factors affecting missing values; such as Establishment type, Area, Legality type and Number of total employees. Expectation Maximization (EM) method was identified as the most appropriate method for missing value imputation in ASI according to the comparison carried out. EM has the lowest mean difference than other four methods.

ASI results provide the annual summary performances of local industries; it also reflects the social and economical state of the country. Hence, accuracy of the ASI results is crucial for making correct policy decisions regar ding the industries. Hence, this study facilitates more reliable data analysis leading to correct decisions for the development of the country.

Keywords: Binary logistic regression model, Expectation maximization, Imputation, Missing values, Survey data