SUMMARY

Missing Value Analysis with Regression and Expectation Maximization (EM) Methods

In various studies, the presence of missing data is a common problem. This could be caused by the respondent deliberately did not answer the question because of an ambiguous question. In addition, the most frequent cause is an error in the implementation of research on interviewer error. In large amount, missing data may effect the results of the research analysis.

Missing value analysis is one of the programs to estimate the missing data. Missing value analysis method consists of listwise deletion, pairwise deletion, regression estimation and expectation maximization (EM). In selecting a suitable method, must be known in advance the pattern value of the available data. If the data types are missing completely at random (MCAR), the method used is the listwise and pairwise deletion. If the data type is missing at random (MAR), estimation method used is regression estimation and the EM.

In addition to errors in the implementation of the study, missing data can also occur in longitudinal studies and experimental research (experimental mortality). Estimation of missing data can be done with the other program than MVA, including the method of maximum likelihood, multiple imputation and generalized estimating equations (GEE).

The purpose of this study was to compare the regression and EM methods in estimating missing data values. This type of research was non-reactive with secondary data analysis. The variables analyzed were age, height and weight of infants in health centers of Bojonegoro regency. Data that was taken consist of 500 infants. The first prosedur was ommitting data with simulation data at 10, 15, and 20% then performed with data imputation with the EM and regression methods to replicate as much as three times. To find the difference of the original data with the results of estimation was tested with the the same subject anova. The best method was determined by looking at the closeness of the highest correlation and the average square of the smallest difference.

Results showed both regression and EM methods no significant differences in mean values and standard deviations. the regression method, a good method was regression with non Adjustment with 2 predictors, the EM method, a good method was EM with 2 predictors and 66.66% for EM methods had on average than the least squares regression methods vary, so it could be interpreted EM method better than the regression method in estimating the missing data.

Regression method was using ordinary least squares approach in which the method of least squares regression coefficient estimator aims at getting the bo and b1 are made of squared errors as small as possible. The weakness in this method lies in the type of data missing at random, making the assessment of b0 and b1 was biased, because the regression modeling based solely on the case of

incomplete data, so that when the type of missing at random, if there was missing data, then the data did not included in model.
EM method based on maximum likelihood approach, namely iterative process, where the initial value was positive so that the value likelihoodnya always rises, until it reached a convergen value.

In general, missing value analysis could estimate the mean, variance and covariance, but could not estimate the value of standard errors. Analysis of maximum likelihood to estimate the standard errors was multiple imputation Conclusion: EM method was better than the regression method in estimating the value of missing data. Nevertheless, the EM method could not estimate the MVA standard errors as in the method of multiple imputation. Therefore, the next researchers need to compare the EM method and multiple imputation methods in determining the best method of estimating missing data.

ABSTRACT

Missing Value Analysis with Regression and Expectation Maximization (EM) Methods

The missing data is the problem which happen in reseaerch that is caused by some factors. In large amount, missing data can influence the validity of research anlysis result. Missing value analysis with regression and EM method is one of methods to estimate missing data. The purpose of this study was to compare the regression and EM methods in estimating missing data values.

This type of research was non-reactive with secondary data analysis. The variables analyzed were age, height and weight of infants in health centers Wisma Indah of Bojonegoro regency. Data that was taken consist of 500 infants. The first prosedur was lossing data with simulation data at 10%, 15%, and 20% then performed with data imputation with the EM and regression methods to replicate as much as three times. To find the difference of the original data with the results of estimation was tested with the the same subject anova. The best method was determined by looking at the closeness of the highest correlation and the average square of the smallest difference.

Results showed both regression and EM methods no significant differences in mean values and standard deviations. the regression method, a good method was regression with non Adjustment with 2 predictors, the EM method, a good method was EM with 2 predictors and 66.66% for EM methods had on average than the least squares regression methods vary, so it could be interpreted EM method better than the regression method in estimating the missing data.

EM method used maximum likelihood approach with iteration process until the value going convergen.

Key word : Regression, EM, Missing data