DOI: 10.17354/ijpphs/2016/27

Original Article

Comparison Between Hotdeck Method and Regression Method in Handling Health Science Missing Data

S K M Onny Priskila[1], M Soenarnatalina[2], N Hari Basuki[2]

[1]Post Graduate Student, Department of Public Health, Faculty of Public Health, Airlangga University, Surabaya, East Java, Indonesia, [2]Lecturer, Department of Statistics, Faculty of Public Health, Airlangga University, Surabaya, East Java, Indonesia

ABSTRACT

Introduction: Missing data or missing value is information that is not available on a subject (case). Missing data occurs because some information on the object is not given, thus it is difficult to find or the actual information does not exist. The case of missing data is ignored as it will certainly make it difficult to obtain a high accuracy for result classification even though the most reliable classification algorithm is used. One method in handling the missing data problem is by imputation. Multiple imputation methods can be used to replace missing data with a constant value, hot deck, regression method, expectation maximization method, and multiple imputation.

Purpose: To analyze, compare, and determine the best imputation method of missing data between hot deck and regression methods.

Materials and Methods: Data used is the data of respondents who practice family planning in the town of Pasuruan, East Java, Indonesia, and age variable. Variable age is used as the simulation data is lost, then imputated by hot deck or regression. The original data results will be compared with the imputed data using *t*-test, Pearson correlation, and root mean square error (RMSE) test.

Results: Results of imputation using simulated data age variable show that regression method is better than hot deck method in handling missing data on health science.

Conclusion: The best method views from the results are not significant *P* value, *r* value close +1, and smallest RMSE value. Hot deck method resulted in *P* value not significant at 5% missing data, but the method has small *r* values even negative and RMSE were great. Regression method resulted in *P* value not significant data missing 5% and 10%. Besides looking at the results of the consistency analysis views also repeat values of *P*, *r,* and RMSE of value three methods.

Key words: Age, Hot deck, Imputation, Missing data, Regression

INTRODUCTION

Missing data or missing value is information that is not available on a subject (case). Missing data occurs because some information on the object is not given, thus it is difficult to find or the actual information does not exist.[1] Based on a mechanism, the type of missing data is divided into three form: Missing completely at random, which is the missing data is not related with the value of all the variables, whether variable with missing data or variable observations. That means missing data is in random. Missing at random (MAR) is the missing data only relates to the response variable/observations. Not MAR, which is the missing data in a variable related to the variable itself, so it cannot be predicted from the other variables in a dataset.[2]

Methods for handling problem of missing data in a statistical analysis are such as procedures based on complete unit (completely recorded units), model-based procedure, weighting procedures, and procedure-based imputation. Multiple imputation methods can be used to replace missing data with a constant value such as hot deck, regression, expectation maximization, and multiple imputation. Some research shows that handling missing data with imputation method can increase classification accuracy than without imputation.[3]

This research will compare two methods of imputation which are hot deck and regression. Hot deck is a complete method of replacing missing data with an average value, especially in prediction standard errors that underestimate.

CORRESPONDING AUTHOR:
S K M Onny Priskila,
JL. Kaliwaron 170 Surabaya, Jawa Timur, Indonesia. Phone: +91-081703011338.
E-mail: onny_priskila@yahoo.com

Submission: 05-2016; Peer Review: 06-2016; Acceptance: 07-2016; Publication: 08-2016

Before using this method, the data must first be sorted by variables assessed variables that are linked to missing data items. People who are in the same cluster are then placed in the same file. The weakness of the hot deck is that the missing data repeatedly filled with value then prediction will be biased.[4] Missing data is obtained by prediction in regression method. Many types of regression models can use in regression imputation for example linear regression and logistic regression. Variable *Y* is obtained from the data with missing data and variable *Z* obtained from the complete data. If *Y* and *Z* are related, then the value *Y* is predictable.

Data that used in this study is by monitoring data of fertile couples with a mini survey of Indonesia in 2014. Mini survey is a research method to collect and analyze a simple data quantitatively and is cheap and fast.[5] Data mini survey on the town of Pasuruan, East Java, Indonesia is used as simulation data for the analysis of missing data. The purpose of the research is to analyze, compare, and determine the best method of imputation of missing data between hot deck and regression methods.

MATERIALS AND METHODS

Type of research is non-reactive research which is a kind of research for secondary data.[6] Data used represents data the respondent town of Pasuruan, East Java, Indonesia, who participated KB. Data amounted to 80 respondents and were taken age variables. The variable age was used as simulation data were then removed as much as 15%, 10%, and 5% by random and repeated three times. Imputation of missing data using methods hot deck and regression, and then the results compared to with the original data imputation. Comparing the results with the original data imputation using *t*-test, Pearson correlation, and root mean square error (RMSE) test.

RESULTS

The data used is variable age of 80 respondents, which is reduced to 15%, 10%, and 5% by random and is repeated three times. In data sets of missing 15%, it is reduced to 12 data if its 10% the data, it is reduced to 8 data and group data of 5% is reduced to 4 data. After the data reduction, empty data is conducted by imputation method hot deck and regression. The reduction of the missing data was repeated 3 times, so imputation is also repeated 3 times. Total 15% missing data produce 36 data, the missing data of 10% total generates 24 data, and the missing data of 5% results in 12 data. Here are the comparison data imputation results with the original data.

From Table 1, it is known that the method missing data hot deck in 15% produce 8.3% data which is same as the original data, the missing data 10% produces 4.1% data which is same as the original data, and the missing data 5% produces 16.6% data also same as the original data. Method of regression on the data missing 15% produces 30.5% data which is same as the original data, the missing data 10% produces 45.8% data which is same data as the original data, and missing data 5% produces 8.3% data which is same data as original data. After imputation, it is then analyzed with paired *t*-test, Pearson correlation, and RMSE.

Table 1: Results imputation of missing data

Method	Data missing		
	15	10	5
Hot deck (%)	8.3	4.1	16.6
Regression (%)	30.5	45.8	8.3

Paired *t*-test

Hypothesis

- H_0: There is no difference between the original data with the data after imputation
- H_1: There is a difference between the original data with the data after imputation.

Results paired *t*-test showed at the missing data 15% has a value significant <α (0.05) is method hot deck imputation at the second imputation *P* = 0.045 and third imputation *P* = 0.034, which means there is a difference between original data with data after imputation. At the missing data 10% has a value significant <α (0.05) is method hot deck imputation at the first imputation P = 0.029, which means there is a difference between original data with data after imputation. At missing data 5% all value not significant, that means there is no difference between original data with data after imputation.

Paired *t*-test is not only seen the value of *P* for each outcome imputation but also patterns of repetition *P* value results. This pattern of results repetition paired *t*-test on the missing data 15% on the method that produces *P* value most stable is hot deck method, whereas missing data 10% and 5% both of methods produces the *P* value is not stable, that means first imputation, second, and third have a result much different *P* values.

Pearson Correlation

Correlation test is used to determine the strong relationship between the original data and data after imputation. If the values are getting closer to r +1, then the relationship is stronger, otherwise if close to −1 then the relationship is getting weaker.

For data missing of 15%, 10%, and 5%, which the value r is close to +1 then regression method is used, which means between the original data and data after imputation with regression methods have a strong relationship. Pearson correlation test is not only judged by the value of *r* of each imputation but also patterns of repetition test results Pearson correlation. From Table 3, it is of Pearson correlation test, the second method that has the most stable value *r* is a regression method.

Table 2: Results paired t-test

Data missing (%)	Hot deck method			Regression method		
	1	2	3	1	2	3
15	*P*=0.102	*P*=0.045	*P*=0.034	*P*=0.764	*P*=1.000	*P*=0.638
10	*P*=0.029	*P*=0.154	*P*=0.662	*P*=0.351	*P*=0.080	*P*=0.195
5	*P*=0.623	*P*=0.282	*P*=0.326	*P*=0.140	*P*=0.058	*P*=0.638

Table 3: Results Pearson correlation

Data missing (%)	Hot deck method			Regression method		
	1	2	3	1	2	3
15	*r*=0.034	*r*=−0.261	*r*=−0.044	*r*=0.085	*r*=0.997	*r*=0.988
10	*r*=0.602	*r*=0.683	*r*=0.332	*r*=0.984	*r*=0.999	*r*=0.953
5	*r*=−0.766	*r*=−0.169	*r*=−0.646	*r*=0.999	*r*=0.999	*r*=0.990

Table 4: Results RMSE

Data missing (%)	Hot deck methods			Regression methods		
	1	2	3	1	2	3
15	7.92	9.17	8.76	8.13	0.58	1.15
10	5.57	7.01	6.62	0.71	0.61	1.84
5	14.15	11.19	11.60	1.58	0.87	0.87

RMSE: Root mean square error

RMSE

The lower RMSE value shows that the variation value produced by a variation of forecast models approached observation. The lower the RMSE value, resulting data is better.

From both of methods that have the smallest RMSE value is regression method. Other than the results of RMSE, RMSE repetition patterns also become one of the considerations in determining the best method. For missing data of 15%, there is no method that has a value is stable, but the method of regression in the second and third imputation has a stable value. At missing data of 10%, there is no method that has RMSE values are stable, but the method of regression in the first and second imputation has a stable value. At missing data of 5% regression method that have the most stable RMSE value, that means the results of imputation first, second, and third resulted RMSE values are not much of a difference.

DISCUSSION

T-test or paired *t*-test is to determine if the samples used have different average or not. *t*-test was used to compare the results imputation data with original data before imputation. *t*-test results were taken if the value is not significant because the data imputation needed was not different from the original data or was close to the original data. Hot deck imputation method produces data which was much larger than the original data, resulting in a larger mean value. This caused a significant *t*-test result. For the regression method, the imputation data was not much of different from the original data.

Correlation or Pearson correlation test is to know the powerful relation between the original data and data after imputation. The overall result of the test Pearson correlation in data group missing 5%, 10%, and 15% showed that regression method produces *r* value closest to + 1, meaning imputation using regression method has a strong relationship between original data with data after imputation. Imputation of missing data with the regression method was obtained by prediction. In this case, the childbirth age variable was used for prediction of the age variable. The age variable and age childbirth variable had a series of data that was almost the same or not much different. This caused the imputation results not very different from original data because the prediction used was not much different, causing the imputation of missing data with the regression method resulted value *r* is the most closer +1, compared to other methods.

For the results of imputation need to be determined the RMSE test to know the results of imputation have large error or not, the smaller value of RMSE the data result is better. RMSE value derived from the square root of the difference between data after imputation with data before imputation, the bigger differences in data before and after imputation the larger is the RMSE value produced and otherwise. This caused regression method have the smallest RMSE values compared to other imputation methods.

CONCLUSION

In conclusion, the best method views from the results are not significant *P* value, *r* value close +1, and smallest RMSE value. Hot deck method resulted in *P* value not significant at 5% missing data, but the method has small *r* values even negative and RMSE were great. Regression method resulted in *P* value not significant data missing 5% and 10%. Besides looking at the results of the consistency analysis views also repeat values of *P*, *r,* and RMSE of value three methods. *t*-test results of the hot deck and regression method resulted in *P* value is stable, whereas at Pearson correlation and RMSE test, regression methods resulted in the most stable patterns. The results of analysis *t*-test, Pearson correlation, repetition and consistency RMSE analysis show that the regression method is better than hot deck method for the analysis of missing data on health science.

REFERENCES

1. Singgih S. Parametric statistics, Concepts and Applications with SPSS. Jakarta: Gramedia; 2010.
2. Rubin DB. Inference and missing data. Biometrika 1976;63:581-92.
3. Farhangfar K. Impact of imputation of missing values on classification error for discrete data. Pattern Recognition 2008;41:3692-705.
4. Ford BL. An overview of hot-deck procedures. In: Madow WG, Olkin I, Rubin DB, editors. Incomplete Data in Sample Surveys, Theory and Bibliographies. Vol. II. New York: AcademicPress:1983. p. 85-207.
5. Wendy H. Monitoring couple fertile age through the Mini SurveyIndonesia. Jakarta: Population and family planning Research, family planning and family National Center welfare; 2013.
6. Kuntoro H. The Philosophical Basic Methodology Research. Surabaya: Pustaka melati; 2011.

HOW TO CITE THIS ARTICLE:
Priskila SKMO, Soenarnatalina M, Basuki NH. Comparison of Imputation Method Hot Deck and Regression to Handling Missing Data on Health Science. Int J Prevent Public Health Sci 2016;2(2):11-13.