



Health Notions

ISSN: 2580-4936
(online version only)

Published by:
Humanistic Network for Science and Technology

Cemara Street 25, RT.01 RW.02, Ds./Kec. Sukorejo, Ponorogo, East Java, Indonesia
63453

Phone: +6285853252665

Email: hunescite@gmail.com, admin@heanoti.com

<http://heanoti.com/index.php/hn>

"Health Notions" is a media for the publication of articles on research, book review and literature review in the areas of health science and practice such as public health, medicine, dentistry, pharmaceutical, environmental health, nursing, midwifery, nutrition, health technology, clinical laboratories, health education, health information system, health management, and health popular.

EDITORIAL TEAM

Editor in Chief:

Dr. Heru Santoso Wahito Nugroho, S.Kep., Ns., M.M.Kes., C.P.M.C.
Leader & Research Consultant, *Forum Ilmiah Kesehatan* (Scientific Forum of Health);
Lecturer, Health Polytechnic of Health Ministry at Surabaya, Indonesia.

Editors:

Dr. David Ackah-Ph.D, CPMC, FPMP, FCISCM, FCICRM, FCE, PGDPM, M.Sc, B.Sc, Dip.
President, Institute of Project Management Professionals, Ghana.

Dr. Michael Burns-Ph.D.
Director, Medical Practitioner, Med Hospital USA

Dr. Hadi Prayitno, Drs., M.Kes.
Lecturer, University of Jember, Indonesia.

Dr. Noer Saudah, S.Kep., Ns., M.Kes.
Lecturer, College of Health Science "Bina Sehat", Indonesia

Dr. Tri Niswati Utami, S.K.M., M.Kes.
Lecturer, Islamic University of Medan, Indonesia.

Dr. Muhammad Anshari, Apt., M.M.
Lecturer, Muhammadiyah University of Banjarmasin, Indonesia.

Dr. Elvis Lotten,
Dean of Medical School, University of Regional Medicals, Costa Rica.

Dr. Huger Vinnit
Head of Pharmacy Department, University of Science and Technology, China.

Wiwin Martiningsih, S.Kep., Ns., M.Kep, Ph.D., NS (c)
Lecturer, Health Polytechnic of Ministry of Health at Malang, Indonesia.

Joel Rey Ugsang Acob, MA, RN, DNS (c)
Lecturer, Faculty of Nursing, Visayas State University, Philippines.

Auta Tanko Titus, RN, Bsc. Nursing Scie., DPA, ADL, PGDE, MSc. Nur Scie. FWACN
Deputy Director, Nursing Services, MOH & Hospital Services, Minna, Niger State,
Nigeria; Lecturer, College of Nursing Sciences, School of Midwifery, Minna, Niger
State, Nigeria

Dr. Yessy Dessy Arna, S.Kp., M.Kep. Sp.Kom.
Lecturer, Health Polytechnic of Health Ministry at Surabaya, Indonesia.

Dr. Indah Lestari, S.Kep., Ns., M.Kes.
Lecturer, College of Health Science "Bina Sehat", Indonesia.

Dr. Kennedy Edem Kukuia
Neuropharmacologist, University of Ghana Hospital; Lecturer, University of Ghana,
Ghana.

Dr. Sahrir Sillehu, S.K.M., M.Kes.
Lecturer, Institute of Health Science "Maluku Husada", Indonesia.

Dr. K. G. Agyenim Boateng
Physician, University of Ghana Hospital; Lecturer, University of Ghana, Ghana.

Dr. Byba Melda Suhita, S.Kep., Ns., M.Kes.
Lecturer, Institute of Health Science "Surya Mitra", Indonesia

Secretariate:

Suparji, S.S.T., S.K.M., M.Pd.
Secretary, *Forum Ilmiah Kesehatan* (Scientific Forum of Health); Lecturer, Health
Polytechnic of Health Ministry at Surabaya, Indonesia.

LIST OF ARTICLES

- **Maternal Access to Health Services by Pregnant Mothers in Bacnotan, La Union, Philippines**
Auta Tanko Titus
151-158
- **"KeIPUS-Gendang Beleq" Model as A Strategy to Improve Knowledge and Attitudes of Couples of Childbearing Age against Long Term Contraception Method**
Awan Dramawan, Moh. Arip, Desty Emilyani
159-162
- **Community Empowerment Model Based on Independence in Administration Alert Village Health Sector**
Suparji Suparji, Heru Santoso Wahito Nugroho, Sunarto Sunarto
163-168
- **Socio-Demographic and Cultural Factors Associated with The Occurrence of Low Birth Weight in Kupang**
Simplicia Maria Anggrahini, Hari Basuki Notobroto, Rika Subarniati Triyoga
169-173
- **Effects of Purple Sweet Potatoes on Oxidative Stress Biomarkers in Rats Subjected to Exhaustive Exercise**
Anindya Maratus` Sholikhah, Bambang Wirjatmadi, Merryana Adriani
174-177
- **The Service Compensation of National Health Insurance and The Performance of Health Workers**
Robiatud Daniyah, Ancah Caesarina Novi Marchianti, Sebastiana Viphindrartin
178-182
- **Comparison of MICE and Regression Imputation for Handling Missing Data**
Berliana Devianti Putri, Hari Basuki Notobroto, Arief Wibowo
183-186
- **Brand Image Effect on Selecting Place of Dentures Fabrication at Integrated Specialist Clinic of RSGM FKG UNAIR**
Ria Nila Bawati, Stefanus Supriyanto, Sri Widati
187-191
- **The Knowledge of Balado Chips Producers on The Use of Dyes in Campago Guguk Bulek Village, Bukittinggi**
Fajrina Hidayati
192-195
- **The Role of Health Workers in Handling People with Mental Disorders (ODGJ) Physical Restraint**
Indriany Maya Ariusta, Sri Hernawati, Isa Ma'rufi
196-200
- **Stimulation Model Growth and Social Personal Development of Children Autism in Health Promotion**
Atik Badi'ah, Ravik Karsidi, Ahmad Arman Subijanto, Diffah Hanim
201-207
- **The Mussel Shell Powder as Natural Preservation for Swordfish by Using Smearing Method**
Demes Nurmayanti, Darjati Darjati, Marlik Marlik
208-211
- **The Comparative Study: Self Acceptance, Environmental Development and Depression on Elderly**
Pipit Festy Wiliyanarti, Irma Rofiatin, Dede Nasrullah, Yulianto Yulianto
212-220
- **Analysis of Risk Factors of Quality of Snacks Food Sold in Town Squares of Magetan, Ngawi, Ponorogo and Madiun**
Denok Indraswati, Djoko Windu P. Irawan, Lilis Prihastini, Sigit Gunawan, Bejo Utomo
221-229



UNIVERSITAS AIRLANGGA

FAKULTAS KESEHATAN MASYARAKAT

Kampus C Mulyorejo Surabaya 601115 Telp.031-5920948, 5920949 Fax 031-5924618

Laman : <http://www.fkm.unair.ac.id>; E-mail: info@fkm.unair.ac.id

SURAT KETERANGAN Nomor : 3148/UN3.1.10/KP/2023

Yang bertandatangan di bawah ini :

Nama : Dr. Santi Martini, dr. M.Kes
NIP : 196609271997022001
Pangkat/Golongan : Pembina / Gol. (IV/a)
Jabatan : Dekan

Dengan ini menerangkan bahwa :

Nama : Dr. Hari Basuki Notobroto, dr., M.Kes
NIP : 196506251992031002
Pangkat/Golongan : Pembina (Gol. IV/a)
Jabatan : Lektor Kepala

Telah melaksanakan penelitian dengan judul sebagai berikut :

No.	Judul Karya Ilmiah	Tahun Pelaksanaan
1	Individual and Neighborhood Socioeconomic status in the prediction of liver transplantation among patients with liver disease (C-18)	2019
2	Survival Analysis of Hemodialysis Patients (C-25)	2016
3	Climate Conditions, Larvae Free Number, DHF Incidence in Surabaya Indonesia (C-37)	2013
4	Development of Air Polluter Model for the Carbon Monoxide (CO) Element Based on Mixed Geographically Temporal Weighted Regression (MGTWR) Kriging (C-38)	2014
5	Comparison Between Hotdeck Method and Regression Method in Handling Health Science Missing Data (C-47)	2016
6	Effect of Climate and Sanitation of Rats Population and Index Flea as Indicator Detection Yersinia Pestis in Port Surabaya (C-51)	2016
7	Factors that Affect the Incidence of HIV/AIDS (An Analysis Using Tobit Regression) (C-57)	2016
8	Demographic Transition and Conditions of Health to Elderly People in East Java Province, Indonesia (C-58)	2016
9	Survival Cox Proportion Hazard in Chronic Kidney Disease with Hemodialysis at Haji General Hospital, Surabaya (C-62)	2018
10	Risk Factors Clustering of Malnutrition Case Based on The Utilization of Toddler Health Service at Lamongan Regency in 2016 (C-63)	2018
11	The Role of the National Health Insurance Program in the Use of Health Services in City X (C-74)	2020
12	Multilevel Analysis of Lifestyle and Household Environment for Toddlers With Symptoms of Acute Respiratory Infection (ARI) in Indonesia in 2007, 2012, and 2017 (C-94)	2022
13	Indikator Karakteristik Fisik Rumah Dominan dalam	2016

	Penentuan Status Kemiskinan untuk Program Rehab Rumah tidak Layak Huni di Kabupaten Sidoarjo (C-110)	
14	Perbandingan Tingkat Konsistensi Normalitas Distribusi Metode Kolmogorov-Smirnov, Lilliefors, Shapiro-Wilk, dan Skewness-Kurtosis (C-105)	2014
15	Penerapan Metode Artificial Neural Network dalam Peramalan Jumlah Kunjungan Ibu Hamil (K4) (C-112)	2019
16	Viability Status of Diabetes Melitus Patients with Complications of Hyperglycaemia, Cetoasidosis, and Gangrene (C-119)	2020
17	Prevalensi Rasio Pelayanan Kesehatan Maternal dan Ketersediaan Fasilitas Kesehatan di ERAJKN/KIS di Indonesia (C-124)	2016
18	Faktor Demografi WUS yang Berhubungan Dengan Status Gizi Berdasarkan LILA di Puskesmas Pacarkeling Tahun 2017 (C-126)	2018
19	Comparison of MICE and Regression Imputation for Handling Missing Data (C-130)	2018
20	The Affecting Factors to Grade of Breast Cancer in Dr. Soetomo Hospital of Surabaya (C-133)	2018
21	Risk Factors of HIV and AIDS Mortality At Ibnu Sina Hospital (C-134)	2018
22	Subset Best Method Regression Analysis with Cp Mallows Statistics on Factors Affecting Life Expectancy (C-135)	2022
23	Hubungan Pola Pengasuhan Orang Tua dengan Tindakan Pencegahan Kekerasan Seksual pada Anak (C-136)	2017

Adapun penelitian tersebut layak dilakukan dan menghasilkan output yang sangat baik, meskipun belum ada *Uji Ethical Clearance* karena merupakan penelitian yang menggunakan data sekunder.

Demikian surat keterangan ini kami buat untuk dapat dipergunakan sebagai persyaratan pengusulan Jabatan Fungsional Guru Besar.

Surabaya, 24 April 2023



Dr. Santi Martini, dr. M.Kes
NIP. 196609271997022001

<http://heanoti.com/index.php/hn>



RESEARCH ARTICLE

URL of this article: <http://heanoti.com/index.php/hn/article/view/hn20207>

Comparison of MICE and Regression Imputation for Handling Missing Data

Berliana Devianti Putri^{1(CA)}, Hari Basuki Notobroto², Arief Wibowo³

^{1(CA)}Faculty of Public Health, Airlangga University, Indonesia; berlianaputri93@gmail.com (Corresponding Author)

²Faculty of Public Health, Airlangga University, Indonesia

³Faculty of Public Health, Airlangga University, Indonesia

ABSTRACT

Data collection activities have a higher risk of missing data. Missing data may produce biased estimates and standard errors increased, so imputation method is needed. The purpose of this study was to investigate which imputation method is the most appropriate to use for handling missing data. The strategies evaluated include complete case analysis, Multivariate Imputation by Chained Equation (MICE), and Regression Imputation. This study was non-reactive study and used raw data RPJMN 2015 Survey from BKKBN East Java Province. There were three incomplete data sets were generated from a complete raw dataset with 5%, 10%, and 15% missing data. Incomplete data sets were made missing completely at random. Based on Friedman Test, both of imputation methods produced estimates which was no different with complete raw data set. Based on Mean Square Error analysis, MICE provided MSE values less and more stable than Regression Imputation in all scenarios. **Conclusion:** Multivariate Imputation by Chained Equation (MICE) was the most recommended method to use for handling missing data less than 15%.

Keywords: Missing data, MICE, Regression imputation

INTRODUCTION

Missing data are a pervasive problem in research using primary data. Missing data can exist due to respondents did not give the answer of questionnaire; negligence of enumerator or data entry; etc. Missing data may increase bias estimates and standard error, so the data set can not be use⁽¹⁾. There are two methods to handle missing data, including case deletion and imputation technique. Case deletion is a classical method that user should remove respondent who had incomplete answer. It would reduce the sample size⁽²⁾.

While imputation technique is a commonly used method to handle missing data. This method is not to deleting respondent, like case deletion, but to predict missing values as close as possible in a way resulting in valid statistical inference. Regression imputation is an imputation technique which replaces missing data with simulated values from predictor, applies standard analyses to each completed dataset, and adjusts the obtained parameter. The lack of this technique is only produce estimates for dependent variables⁽²⁾.

Recently, Multivariate Imputation by Chained Equation (MICE) is an innovation of imputation method which has been completed by chained equation, so it can be recommended way for handle missing data. Chained equation principled on *Markov Chain Monte Carlo* (MCMC). It makes this method be more flexible than others. It can produce estimates for not only dependent variable but also independent variables at the same time^(3,4).

The purpose of this study was to investigate which imputation method is the most appropriate to use for handling missing data, whether Multivariate Imputation by Chained Equation (MICE) or Regression Imputation.

METHODS

This study was non-reactive study and used raw data RPJMN 2015 Survey from BKKBN East Java Province⁽⁵⁾. The population was all respondents of RPJMN 2015 Survey which had dating and pre-marital sexual intercourse. This study used six variables, including age of first puberty, age of first dating, knowledge of contraceptives, knowledge of HIV/AIDS, knowledge of fertile period, and age of first sexual intercourse. The dependent variable of this study was age of first sexual intercourse. The variables that would be used as a simulation variables were age of first dating and age of first sexual intercourse. Simulation data sets were artificially made MAR (Missing at Random). Here were the regression imputation steps using STATA software:

1. Save the simulation data set with file extension *.dta
2. Open the simulation data set using command `use name_of_dataset`
3. Establish `dataset mi` using command `mi set wide`

4. Choose the variable that will be imputed using command **mi register imputed name_of_variable**
5. Choose the variables that will be predictor using command **mi register regular name_of_variable**
6. Ensure the amount of missing data using command **misstable sum**
7. Imputation using command **mi impute regress name_of_imputed (space) name_of_predictor, add(5) rseed(1500)**
8. Incomplete data set has been completed with imputation values.

Regression imputation could be able to replace missing values with only one predictor variable. Before do the regression imputation, we need to know what variable is appropriate to be a predictor and put the missing variable as dependent variable. If the age of first sexual intercourse were imputed, the age of first dating would become a predictor variable. If the age of first dating were imputed, the age of first puberty would become a predictor variable. Whereas the Multivariate Imputation by Chained Equation (MICE) could be able to replace missing values all of variables at the same time. Here were the MICE steps using STATA software:

1. Save the simulation data set with file extension ***.dta**
2. Open the simulation data set using command **use name_of_dataset**
3. Establish *dataset mi* using command **mi set wide**
4. Choose the variable that will be imputed using command **mi register imputed name_of_variables**
5. Choose the variables that will be predictor using command **mi register regular name_of_variables**
6. Ensure the amount of missing data using command **misstable sum**
7. Imputation using command **mi impute chained (pmm, knn(5)) name_of_imputed1 name_of_imputed2 = name_of_predictor1 name_of_predictor2 name_of_predictor3 name_of_predictor4, add(5) rseed(1500)**
8. Incomplete data set has been completed with imputation values

Above were the regression imputation steps and the Multivariate Imputation by Chained Equation (MICE) steps^{(3,4),(6,7)}. Incomplete data sets were analysed three times to see their stabilization during produce estimate values. The strategies evaluated include complete case analysis, MICE, and Regression imputation. The comparative parameter were Mean Square Error (MSE) values and Friedman test.

RESULTS

The total sample size was 1.646 respondents. There were three incomplete data set were generated from a complete raw data set with 5% ($n_{\text{missing}} = 83$), 10% ($n_{\text{missing}} = 165$), and 15% ($n_{\text{missing}} = 247$). Here were the imputation results with five iterations, continued with comparative test using Friedman Test (Table 1, 2, and 3).

Table 1. Comparison of MICE and Regression Imputation on 5% Missing Data

Variables	\bar{X}_{ori}	Regression Imputation				MICE				p-value*
		\bar{X}_{imp}	Ties		\bar{X}_{imp}	Ties				
			n	%		n	%			
Age of first sexual intercourse										
1 st analysis	17.63	17.31	33	39.76%	17.25	19	22.89%	0.255		
2 nd analysis	17.69	17.76	4	4.82%	17.59	10	12.05%	0.491		
3 rd analysis	17.61	17.62	2	2.41%	17.50	3	3.61%	0.476		
Age of first dating										
1 st analysis	15.49	15.43	26	31.33%	15.83	16	19.28%	0.475		
2 nd analysis	15.90	15.52	0	0.00%	15.65	3	3.61%	0.321		
3 rd analysis	15.87	15.42	0	0.00%	15.58	4	4.82%	0.198		

*Friedman Test ($\alpha=0,05$)

Table 2. Comparison of MICE and Regression Imputation on 10% Missing Data

Variables	\bar{X}_{ori}	Regression Imputation				MICE				p-value*
		\bar{X}_{imp}	Ties		\bar{X}_{imp}	Ties				
			n	%		n	%			
Age of first sexual intercourse										
1 st analysis	17.84	17.70	28	16.97%	17.79	41	24.85%	0.755		
2 nd analysis	17.62	17.51	4	2.42%	17.52	4	2.42%	0.994		
3 rd analysis	17.77	17.61	36	21.82%	17.57	3	1.82%	0.553		
Age of first dating										
1 st analysis	15.67	15.65	28	16.97%	15.73	32	19.39%	0.346		
2 nd analysis	15.36	15.59	0	0.00%	15.38	36	21.82%	0.249		
3 rd analysis	15.63	15.58	2	1.21%	15.57	5	3.03%	0.285		

*Friedman Test ($\alpha=0,05$)

Table 3. Comparison of MICE and Regression Imputation on 15% Missing Data

Variables	\bar{X}_{ori}	Regression Imputation			MICE			<i>p-value*</i>
		\bar{X}_{imp}	Ties		\bar{X}_{imp}	Ties		
			n	%		n	%	
Age of first sexual intercourse								
1 st analysis	17.66	17.70	46	18.62%	17.49	43	17.41%	0.216
2 nd analysis	17.62	17.67	12	4.86%	17.78	12	4.86%	0.051
3 rd analysis	17.55	17.59	15	6.07%	17.60	0	0.00%	0.057
Age of first dating								
1 st analysis	15.65	15.65	47	19.03%	15.66	44	17.81%	0.759
2 nd analysis	15.64	15.62	1	0.40%	15.41	10	4.05%	0.118
3 rd analysis	15.36	15.70	2	0.81%	15.57	15	6.07%	0.061

*Friedman Test ($\alpha=0,05$)

Table 4. Comparison of MICE and Regression Imputation Based on Mean Square Error

Percentage	Variables	MSE values						Best Method
		Regression Imp.			MICE			
		1	2	3	1	2	3	
5%	Age of first sexual intercourse	3.25	3.23	4.59	2.81	3.13	4.46	MICE
	Age of first dating	5.08	3.89	5.55	5.04	2.79	3.20	
10%	Age of first sexual intercourse	4.21	4.45	4.34	4.35	4.16	4.30	MICE
	Age of first dating	4.28	4.75	4.10	3.36	3.70	2.99	
15%	Age of first sexual intercourse	4.67	4.65	3.33	4.50	4.52	3.22	MICE
	Age of first dating	4.71	5.06	4.28	4.23	3.69	3.91	

Incomplete data sets were analysed three times to see their stabilization during produce estimate values. As shown in Table 1, Table 2, and Table 3, MICE and Regression Imputation produced estimate values similar to the original data set. Both of them had mean as close as original data set. Then, the friedman test showed that there was no different between MICE data set, Regression Imputation data set, and original data set (p -value < 0.05). It also showed that MICE method was more stable containing ties than Regression Imputation in all scenarios. Table 4 showed that MICE produced MSE less than Regression Imputation in all scenarios.

DISCUSSION

Incomplete data sets were artificially made MAR (Missing at Random) so that the estimate values was not biased⁽⁷⁾. User could using Little’s MCAR Test to determine whether incomplete data sets were random or not. If the significant values more 0.05, it would be random. The data sets of this study had significant values more than 0.05, so that imputation technique could be done on all of data sets.

This study used MICE method and Regression Imputation method with five times iteration for imputing missing data. The principle of Regression Imputation was fill in the missing values by using one of the most influences its variable⁽⁸⁾. It caused Regression Imputation should through correlation analysis first to determine the predictor. While the MICE method did not need to do that, because it already completed by MCMC, predictive mean matching, and k-nearest neighbours^{(3),(6)}.

There were two imputation data sets to be compared with original data set using Friedman Test. It showed that there was no different between MICE data set, Regression Imputation data set, and original data set. This result was similar to research conducted by Rakhmat (2010) which showed that Regression Imputation and Predictive Mean Matching Imputation could predict missing values as close as original values. Both of them involved other variable as a predictor, so that imputation values similar to the original values. It was seen from the number of ties in each imputation methods⁽⁹⁾.

Then, one of the comparison parameters to know the accuracy of imputation results was Mean Square Error (MSE). The principle of MSE was to calculate the differences between original values and the imputation values. An imputation method which produced the smallest MSE was the best imputation method for handle missing data because it predicted missing values as close as possible to the true ones in a way resulting in valid statistical inference.

The MSE analysis showed that MICE produced MSE values less than Regression Imputation, either on 5% missing data, 10% missing data, or 15% missing data. It could be concluded that Multivariate Imputation by Chained Equation was better to use for handling missing data than Regression Imputation. MICE was an innovation of imputation method which completed by Markov Chained Monte Carlo (MCMC), Predictive Mean Matching (PMM), K-Nearest Neighbours (KNN). It could be seen in the command written in the MICE steps.

Syntax pmm, knn(5) could produce estimated values from the new regression model with considering other units. It caused MICE method could predict missing values as close as original values^{(4),(9)}.

CONCLUSION

Multivariate Imputation by Chained Equation (MICE) was the most recommended method to use for handling missing data less than 15%. Hopefully, this result could help other researchers to handle missing data with not to deleting unit which had missing value but using MICE method to complete all values.

REFERENCES

1. Farhangfari A, Kurgan L, Dy J. Impact of Imputation of Missing Values on Classification Error for Discrete Data. *Pattern Recognition*. 2008;41(12):3692-3705
2. Yuan YC. *Multiple Imputation for Missing Data: Concept and New Development (Version 9.0)*. SAS Institute Inc; 2001.
3. Royston P, White IR. Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *Journal of Statistical Software*. 2011;45(4):45-57.
4. Buuren S. *Flexible Imputation of Missing Data*. CRC Press; 2012.
5. BKKBN. *Results of RPJMN 2015 Survey (Hasil Survei RPJMN Tahun 2015 Publikasi)*. Jakarta: Puslitbang BKKBN; 2015.
6. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. *Multiple Imputation by Chained Equations: What is it and How does it Work?* *International Journal of Methods in Psychiatric Research*. 2011;20(1):40-49.
7. Harlan J. *Missing Data and Multiple Imputation (Data Kosong dan Imputasi Ganda)*. Jakarta: Penerbit Universitas Gunadarma; 2016.
8. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. John Wiley & Sons Publishing; 2012.
9. Rakhmat B. *Multiple Imputation Using Regression and Predictive Mean Matching for Imputing Missing Data (Imputasi Berganda Menggunakan Metode Regresi dan Metode Predictive Mean Matching untuk Menangani Missing Data)*. Master Thesis. Surabaya: Institut Teknologi Sepuluh Nopember; 2010.