Jurnal EurekaMatika, Vol. 9, No. 1, 2021
Journal homepage: https://ejournal.upi.edu/index.php/JEM
e-ISSN: 2528-4231, p-ISSN: 2776-480X

Estimasi *Missing Data* dengan Metode *Multivariate Imputation by Chained Equations* (Mice) untuk Membentuk Persamaan Regresi Linear Berganda

Irma Eldiyana[1)], Elah Nurlaelah[2)], Nar Herrhyanto[2)]

[1)]Mahasiswa Departemen Pendidikan Matematika FPMIPA UPI
[2)]Dosen Departemen Pendidikan Matematika FPMIPA UPI
*Surel: irmaeldiyana@gmail.com

ABSTRAK Penelitian ini bertujuan untuk menentukan mengestimasi data yang kosong atau data yang hilang *(Mising data)*. *Missing data* adalah hilangnya sebagian informasi atau sebagian data pada suatu penelitian. Metode yang digunakan untuk mengatasi *missing data* pada artikel ini *Multivariate Imputation by Chained Equation* (MICE). Penerapan MICE terdiri dari tiga langkah utama, yaitu imputasi, analisis, dan *pooling*. Hasil analisis terhadap data sekunder menghasilkan diperlukan lima kali imputasi untuk mengisi *missing data*. Langkah analisis menggunakan analisis regresi linear berganda, dengan lima model fit. Kemudian pada langkah *pooling*, ke-lima model fit regresi linear berganda yang dihasilkan digabungkan menjadi model pool. Selanjutnya model pool yang diperoleh dibandingkan dengan model regresi berganda data awal. Hasil perbandingan menunjukkan bahwa persamaan linear berganda dengan *missing data* yang diestimasi metode MICE mendekati persamaan liner berganda yang disusun dari data awal, dengan demikian estimasi *missing data* dengan metode MICE dapat dikatakan baik untuk digunakan.

Kata Kunci: *Missing data, Multivariate Imputation by Chained Equation* (MICE), persamaan regresi linear berganda

e-ISSN: 2528-4231, p-ISSN: 2776-480X

Estimation of Missing Data Using Multivariate Imputation by Chained Equations Method to Form Multiple Linear Regression Equations

ABSTRACT *This research aims to determine the value of missing data. Missing data is the loss of some information or data in a research. The method was used to overcome this missing data was the Multivariate Imputation by Chained Equation (MICE). Implementation of MICE in this article contained three main steps, there were imputation, analysis and pooling. Based on analysis on data, it needs five steps of the imputation to fill in the missing data. The analysis steps used the multiple regression analysis with five fit models. And then in the pooling step, the five fit models were combined into one to obtain a pool model. The resulted pool model will be compared with the regression model with the initial data using Root Mean Square Error (RMSE). By looking at the results of the RMSE, the MICE method is considered to be appropriate to use for estimating missing data.*

Keywords: *Missing data, Multivariate Imputation by Chained Equation (MICE), multiple linear regression equations*

1. PENDAHULUAN

Pada tahap pengumpulkan data seringkali terjadi hambatan, salah satu hambatannya yakni fenomena *missing data* atau data hilang. *Missing data* adalah hilangnya sebagian informasi atau data pada suatu penelitian. Permasalahan *missing data* biasa ditemui di berbagai bidang, namun hampir pada setiap pelaksanaan survei atau sensus, terdapat variabel yang non-respon (Rubin, 1987). Beberapa hal yang menyebabkan *missing data,* misalnya dari peralatan yang tidak berfungsi dengan baik, kekurangan fasilitas, tidak terisinya kuesioner karena penolakan responden atau responden kesulitan untuk menjawab pertanyaan, kesalahan dalam pengambilan data, dan lain sebagainya.

Akibat dari adanya *missing data* adalah pendugaan parameter menjadi tidak efisien. Ukuran data yang berkurang dapat mengakibatkan kesulitan dalam menganalisis, sehingga hasil yang didapatkan menjadi tidak valid dan tujuan dari penelitian tidak tercapai. *Missing data* dapat saja diabaikan, jika data yang hilang sedikit. Namun apabila *missing data* berjumlah cukup besar maka data tersebut tidak dapat diabaikan. Oleh karena itu, perlu dilakukan estimasi untuk mengisi non respon tersebut agar inferensi statistik untuk data lengkap dapat dilakukan (Rubin, 1987).

Terdapat beberapa metode yang dapat digunakan untuk mengestimasi nilai dari *missing data* tersebut yang dikelompokkan menjadi dua, yaitu metode tradisional dan metode modern. Metode modern muncul karena keterbatasan dari metode tradisional. Salah satu dari metode modern, yaitu metode *Multivariate Imputation by Chained Equations* (MICE). MICE dikenal juga dengan *"Fully Conditional Spesification"* atau *"Sequential Regression Multiple Imputation"* yang telah muncul dalam literatur statistika sebagai salah satu metode prinsip untuk menangani *missing data* (Azur et al., 2011).

MICE dapat digunakan untuk berbagai model data, seperti data kontinu, data biner (regresi logistik), data kontinu 2-level, regresi logistik polikotomus, dan odds proporsional (Azur et al., 2011), (Zhang, 2016). Prosedur MICE mengikuti serangkaian model regresi yang dijalankan, dimana masing-masing variabel dari data yang hilang dimodelkan bersyarat pada variabel lain dalam data tersebut. Ini berarti bahwa setiap variabel dapat dimodelkan menurut distribusinya (Wulff & Jeppesen, 2017), (Resche-Rigon & White, 2018).

Terdapat beberapa penelitian sebelumnya terkait dengan penggunaan MICE untuk mengestimasi data hilang. Bouhlila & Sellaouti (2013) menggunakan MICE untuk memperkirakan missing data pada TIMSS *database*. Ferguson et al. (2018) membandingkan pendekatan model campuran linier parametrik (LMM) dengan beberapa imputasi dengan MICE untuk studi epidemiologi faktor risiko pada

pembatasan pertumbuhan. Roystaon & White (2011) menggunakan MICE dengan data dari hasil observasi dari kanker ovarium untuk mengilustrasikan variable yang paling penting dati beberapa pilihan yang tersedia. Shah et al. (2014) membandingkan MICE parametrik dengan algoritma MICE berbasis hutan acak dalam dua studi simulasi.

Pada artikel ini, data lengkap yang diperoleh akan dihilangkan beberapa buah dengan mengikuti aturan tertentu, kemudian data yang hilang (*missing data)* itu akan diestimasi dengan MICE. Selanjutnya akan dibentuk persamaan regresi linear berganda yang berasal dari data awal dan data hasil estimasi *missing data*. Persamaan regresi yang terbentuk akan dibandingkan untuk melihat bagaimana efektivitas penggunaan metode MICE dalam mengatasi *missing data.*

2. METODE PENELITIAN

Metode yang digunakan pada penelitian ini adalah deskriptif kuantitatif, karena data yang digunakan dalam penelitian ini berupa data sekunder, yaitu data pengukuran bagian badan 33 orang polisi wanita di daerah Amerika dan Afrika (Van Buuren & Groothuis-Oudshoorn, 2011). Variabel-variabel yang diukur adalah *sitting height* (SITHT), *upper arm length* (UARM), *hand width* (HAND), *upper leg length* (ULEG), *lower leg length* (LLEG), dan *foot length* (FOOT) dan *high* (tinggi badan). Selanjutnya dari data yang diperoleh akan disusun persamaan regresi linier berganda dengan tinggi badan sebagai variabel tidak bebas dan variabel lainnya sebagai variabel bebas.

Selanjutnta data awal tersebut akan disimulasikan sehingga mengandung *missing data*, kemudian *missing data* tersebut diestimasi menggunakan metode *Multivariate Imputation by Chained Equations* (MICE). MICE merupakan bagian dari metode imputasi ganda. Imputasi ganda adalah metode pilihan untuk masalah data yang tidak lengkap dalam kondisi yang kompleks (Van Buuren & Groothuis-Oudshoorn, 2011). Kondisi yang kompleks disini adalah ketika terdapat *missing data* yang lebih dari satu variabel.

Prosedur yang dilakukan dalam penelitian ini adalah:

1. Mempersiapkan data lengkap (data awal sebelum dilakukan pembentukan *missing data*).
2. Melakukan analisis regresi linear berganda pada data awal yang lengkap.
3. Pembentukan *missing data* dengan persentase *missing data* sebesar dua persentase, yaitu pada persentase 15% dan 30%. Penghapusan dilakukan

secara acak dengan tipe *missing data,* yakni MAR dan pola data hilang nonmonoton yakni pola multivariat.
4. Melakukan tiga langkah utama (imputasi, analisis, pooling) pada data yang mengandung *missing data* sehingga terbentuk set data lengkap hasil imputasi.
5. Melakukan analisis regresi linear berganda pada data lengkap hasil imputasi.
6. Membandingkan hasil analisis regresi linear berganda pada data awal dan data hasil imputasi berdasarkan hasil persamaan regresinya, juga dilihat hasil RMSE untuk mengetahui seberapa baik metode MICE untuk estimasi *missing data*.

3. TEMUAN DAN PEMBAHASAN

Berdasarkan langkah-langkah yang diuraikan pada prosedur penelitian di atas maka langkah pertama yang harus dilakukan adalah menyusu persamaan analisis regresi linear berganda dari data awal. Sebelum memperoleh persamaan regresi tersebut, dilakukan pengujian asumsi normalitas, dan hasil analisis menunjukkan bahwa data berdistribusi normal. Setelah itu diperoleh persamaan regresi linear berganda dengan data awal sebagai berikut:

$$\hat{Y} = 15.932 + 0.877X_1 - 0.280X_2 + 0.941X_3 + 0.166X_4 + 1.278X_5 + 1.056X_6$$

3.1 Analisis Data untuk Estimasi *Missing Data*

Tiga langkah utama dalam MICE, adalah langkah imputasi, langkah analisis, dan langkah pooling (Resche-Rigon & White, 2018). Pengolahan data dilakukan dengan bantuan software R-Studio dan menggunakan *package mice*.

Data awal yang diperoleh dari data sekunder dihilangkan (*missed*) sebanyak 15% dan 30%. Selanjutnya data yang hilang (*missing data)* tersebut akan diestimasi lagi menggunakan metode MICE. Melalui langkah imputasi ini *missing data* akan diestimasi dengan nilai yang diperkirakan cukup layak, selanjutnya dibandingkan data hasil estimasi dengan data awal. Teknik imputasi muncul untuk memperbaiki teknik sebelumnya yang sering kurang tepat.

Proses imputasi ini dilakukan sebanyak 5 kali iterasi terhadap masing-masing variabel yang mengandung *missing data* (Van Buuren & Groothuis-Oudshoorn,

2011). Dalam hal ini proses membangun imputasi diawali dengan memanggil fungsi *mice* dan diperoleh hasil imputasi *missing data* sebagai berikut:

1. Hasil imputasi *missing data* 15%

```
> imp$imp$UARM
      1    2    3    4    5
3  35.0 33.2 32.5 33.6 33.2
12 32.6 32.3 33.6 32.5 32.3

> imp$imp$ULEG
     1  2    3    4  5
7 40.3 42 38.6 40.1 41

> imp$imp$FOOT
     1   2   3   4   5
18 4.9 5.2 5.1 5.9 4.9
29 5.2 5.2 5.1 5.9 4.9
```

2. Hasil imputasi *missing data* 30%

```
> imp$imp$SITHT
      1    2    3    4    5
21 88.2 87.1 89.6 88.7 87.1
30 83.7 84.9 83.9 83.9 84.9

> imp$imp$UARM
      1    2    3    4    5
3  36.2 35.0 32.8 32.8 33.2
15 30.5 32.8 32.3 31.0 31.0
25 32.5 33.5 33.2 33.2 31.5

> imp$imp$HAND
      1    2    3    4    5
12 20.2 18.2 18.3 20.2 19.8
33 19.1 18.2 19.1 18.3 18.7

> imp$imp$ULEG
     1    2    3  4    5
7 44.2 38.6 40.3 41 38.9
```

```
> imp$imp$FOOT
     1   2   3   4   5
18 4.9 5.6 4.9 5.9 6.1
29 4.9 5.7 5.7 5.9 5.2
```

Berikut diberikan tabel hasil data imputasi yang diperoleh dengan metode MICE:

1. Imputasi untuk *missing data* 15%

   Pada Tabel 1 disajikan data awal yang dihilangkan sebanyak 15% dari seluruh data yang ada. Selanjutnya *missing data* diestimasi dengan metode MICE sehingga diperoleh data hasil imputasi sebagai berikut:

Tabel 1. Data Awal dan Data Hasil Imputasi 15%

Nomor/Variabel	Data Awal yang Dihilangkan	Data Hasil Imputasi 15%	Residual
3/UARM	33,6	35	-1,4
12/UARM	34,3	32,6	1,7
7/ULEG	40	40,3	-0,3
18/FOOT	5	4,9	0,1
29/FOOT	6	5,2	0,8

2. Imputasi untuk *missing data* 30%

   Pada Tabel 2 disajikan data awal yang dihilangkan sebanyak 30% dari seluruh data yang ada. Selanjutnya *missing data* diestimasi dengan metode MICE sehingga diperoleh data hasil imputasi sebagai berikut;

2. *Data Awal dan Data Hasil Imputasi 30%*

Nomor/Variabel	Data Awal yang Dihilangkan	Data Hasil Imputasi 30%	Residual
21/SITHT	88,1	88,2	-0,1
30/SITHT	85	83,7	1,3
3/UARM	33,6	36,2	-2,6
15/UARM	30,6	30,5	0,1
25/UARM	35,2	32,5	2,7
12/HAND	19,2	20,2	-1
33/HAND	19,4	19,1	0,3
7/ULEG	40	44,2	-4,2
18/FOOT	5	4,9	0,1
29/FOOT	6	4,9	1,1

Setelah dilakukan imputasi terhadap *missing data* tersebut, data yang tidak lengkap menjadi lengkap kembali. Data yang diperoleh disimpan dalam kelas *mids*. Selanjutnya disusun analisis regresi linear berganda, utntu melihat kesesuaian persamaan analisis berganda dilakukan analisis model fit. Hasil dari analisis ini diperoleh lima model fit. Proses analisis model fit ini dilakukan dengan fungsi *with.mids()* dan disimpan dalam kelas *mira.* Kelima model fit memberikan estimasi model regresi yang tidak jauh berbeda satu dengan yang lainnya.

Langkah terakhir adalah langkah pooling yakni menentukan model regresi akhir, selanjutnya dari kelima model fit diambil satu model yang dinamakan model pool. Proses ini dikerjakan dengan fungsi *pool()*, sedangkan hasilnya disimpan dalam kelas *mipo.*

1. Analisis Model Pool untuk *missing data* 15%

Tabel 3. Output analisis model Pool untuk *missing data* 15%

Class: mipo m = 5	Estimate
(Intercept)	16.8307981
SITHT	0.8817514
UARM	-0.3597665
HAND	0.9335845
ULEG	0.1687326
LLEG	1.3163289
FOOT	1.0145962

Berdasarkan output pada Table 3 diperoleh persamaan regresi linear berganda dari model pool yaitu:

$$\hat{Y} = 16.831 + 0.882X_1 - 0.360X_2 + 0.934X_3 + 0.169X_4 + 1.316X_5 + 1.015X_6$$

2. Analisis Model Pool untuk *missing data* 30%

Tabel 4. Output analisis model Pool untuk *missing data* 30%

Class: mipo m = 5	Estimate
(Intercept)	18.2463608
SITHT	0.8581877
UARM	-0.3999731
HAND	0.9798361
ULEG	0.1879306
LLEG	1.3319715
FOOT	0.9506686

Berdasarkan output pada Tabel 4 diperoleh persamaan regresi linear berganda dari model pool yaitu:

$$\hat{Y} = 18.246 + 0.858X_1 - 0.399X_2 + 0.979X_3 + 0.188X_4 + 1.332X_5 + 0.951X_6$$

3.2 Persamaan Regresi Data Awal Dengan Data Hasil Imputasi

1. Persamaan regresi linear berganda dengan data awal lengkap

   $\hat{Y} = 15.932 + 0.877X_1 - 0.280X_2 + 0.941X_3 + 0.166X_4 + 1.278X_5 + 1.056X_6$

   Hasil $Adjusted\ R^2$ sebesar 0,862 atau sama dengan 86,2% .

2. Persamaan regresi linear berganda dengan data hasil imputasi *missing data* 15%

   $\hat{Y} = 16.831 + 0.882X_1 - 0.360X_2 + 0.934X_3 + 0.169X_4 + 1.316X_5 + 1.015X_6$

   Hasil $Adjusted\ R^2$ sebesar 0,865 atau sama dengan 86,5%.

3. Persamaan regresi linear berganda dengan data hasil imputasi *missing data* 30%

   $\hat{Y} = 18.246 + 0.858X_1 - 0.399X_2 + 0.979X_3 + 0.188X_4 + 1.332X_5 + 0.951X_6$

   Hasil $Adjusted\ R^2$ sebesar 0,866 atau sama dengan 86,6% .

3.3 Perbandingan Persamaan Regresi Data Awal Dengan Data Hasil Imputasi

Setelah diperoleh tiga hasil persamaan regresi pada bagian sebelumnya, dapat dilihat bahwa koefisien dari persamaan-persamaan regresi tersebut tidak jauh berbeda. Untuk memperkuat asumsi bahwa persamaan regresi yang diperoleh dengan data hasil imputasi tersebut baik untuk digunakan, akan dilihat juga hasil perbandingan nilai RMSE yang diperoleh sebagai berikut:

1. Nilai RMSE dari persamaan regresi dengan data hasil imputasi *missing data* 15% sebesar 0,207
2. Nilai RMSE dari persamaan regresi dengan data hasil imputasi *missing data* 30% sebesar 0,438

Dari hasil perhitungan di atas, dapat dilihat bahwa hasil RMSE yang lebih kecil yaitu persamaan regresi dengan data hasil imputasi dengan persentase 15%. Berdasarkan teorinya dikatakan bahwa semakin kecil nilai RMSE yang dihasilkan, semakin bagus pula hasil estimasi yang dilakukan (Van Buuren & Groothuis-Oudshoorn, 2011), sehingga dapat dikatakan bahwa data hasil imputasi *missing data* menggunakan metode MICE cukup baik untuk digunakan dan dianalisis dengan regresi linear berganda.

4. KESIMPULAN

Imputasi data MICE digunakan untuk menangani masalah *missing data* atau data hilang. *Missing data* dilakukan secara acak sebanyak 15% dan 30% dengan tipe *missing data* MAR.

Langkah analisis model fit menghasilkan lima hasil persamaan regresi linear berganda yang selanjutkan digabung menjadimodel pool, sehingga terbentuk satu persamaan regresi untuk data dengan *missing data.* Persamaan linear yang terbentuk tidak jauh berbeda dengan persamaan regresi dengan data awal. Persamaan regersi untuk masing-masing dapat dilihat sebagai berikut:

a. Persamaan regresi linear berdasarkan data awal yakni:

$$Y = 15.932 + 0.877X_1 - 0.280X_2 + 0.941X_3 + 0.166X_4 + 1.278X_5 + 1.056X_6.$$

b. Persamaan regresi linear berdasarkan data hasil imputasi *missing data* 15%, yakni:

$$Y = 16.831 + 0.882X_1 - 0.360X_2 + 0.934X_3 + 0.169X_4 + 1.316X_5 + 1.015X_6$$

dengan RMSE sebesar 0,207.

c. Persamaan regresi linear berdasarkan data hasil imputasi *missing data 30*%, yakni:

$$Y = 18.246 + 0.858X_1 - 0.399X_2 + 0.979X_3 + 0.188X_4 + 1.332X_5 + 0.951X_6$$

dengan hasil RMSE sebesar 0,438.

Berdasarkan hasil RMSE dapat dilihat bahwa nilai RMSE paling kecil terdapat pada persamaan regresi dengan data hasil imputasi dengan persentase 15%,

sehingga metode MICE untuk estimasi *missing data* untuk data tersebut dikatakan cukup baik.

5. DAFTAR PUSTAKA

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International journal of methods in psychiatric research*, *20*(1), 40-49.

Bouhlila, D. S., & Sellaouti, F. (2013). Multiple imputation using chained equations for missing data in TIMSS: a case study. *Large-scale Assessments in Education*, *1*(1), 1-33.

Ferguson, K. K., Yu, Y., Cantonwine, D. E., McElrath, T. F., Meeker, J. D., & Mukherjee, B. (2018). Foetal ultrasound measurement imputations based on growth curves versus multiple imputation chained equation (MICE). *Paediatric and perinatal epidemiology*, *32*(5), 469-473.

Resche-Rigon, M., & White, I. R. (2018). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical methods in medical research*, *27*(6), 1634-1649.

Royston, P., & White, I. R. (2011). Multiple imputation by chained equations (MICE): implementation in Stata. *Journal of statistical software*, *45*, 1-20.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Canada: Library of Congress Cataloging in Publication.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American journal of epidemiology*, *179*(6), 764-774.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, *45*, 1-67.

Wulff, J. N., & Jeppesen, L. E. (2017). Multiple imputation by chained equations in praxis: guidelines and review. *Electronic Journal of Business Research Methods*, *15*(1), 41-56.

Zhang, Z. (2016). Multiple imputation with multivariate imputation by chained equation (MICE) package. *Annals of translational medicine*, *4*(2).