CAUCHY: Jurnal Matematika Murni dan Aplikasi
Volume 9(1) (2024), Pages 94-105
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Estimating missing panel data with regression and multivariate imputation by chained equations (MICE)

Budi Susetyo[1*], Anwar Fitrianto[1,2]

[1] Department of Statistics IPB University, Indonesia
[2] Institute of Engineering Mathematics, Universiti Malaysia Perlis, Malaysia

Email: budisu@apps.ipb.ac.id

ABSTRACT

Missing data may occur in various types of research. Regression and multiple imputation by chained equations (MICE) are two methods that can be used to estimate missing data in panel data types. This study aims to compare the accuracy of the missing panel data estimation using the regression and the MICE methods. The data used in this study are 161 random samples of senior high schools and vocational schools in DKI province for the year 2016-2020. Based on the results of the Chow test, Hausman test, and Lagrange Multiplier test on panel data regression, it shows that the appropriate model for the student-teacher ratio (X5) is random, the percentage of teachers who have an educator certificate (X6) is a fixed model with the specific effect of individual school and time, while the percentage of teachers who hold a bachelor degree (X7) is a fixed model with the specific effect of individual. Based on this model, the estimation of missing data is then carried out. The accuracy of the missing data estimation was carried out by comparing the MAPE, MAE, and RMSE values. The results show that the MICE method is quite good for estimating missing data at X5, quite feasible for estimating X6, and very good for estimating missing data at X7. In general, MICE is more accurate than panel data regression.

Keywords: Missing data, Panel data, Imputation, Regression, Multiple imputation by chained equations.

INTRODUCTION

Panel data combines cross-sectional and time series data structures consisting of a set of observation units and variables collected over several periods [1], [2] . Panel data is also known as longitudinal data which assumes that the observations of a unit of observation are not mutually exclusive between periods. Therefore, special models and methods are needed to analyze panel data, [3]. The basic education data that is annualy collected by the Ministry of Education, Culture, Research, and Technology (often known as DAPODIK) is an example of panel data. The DAPODIK annually collects data on the development of schools, students, teachers, assessment results, school facilities, and infrastructure.

Various problems may appear in dealing with longitudinal education data, such as the presence of outliers and missing data. Missing data is a situation where there are

Submitted: December 13, 2023 Reviewed: December 16, 2023 Accepted: January 29, 2024
DOI: http://dx.doi.org/10.18860/ca.v9i1.24824

observations in the dataset that are empty or have no value, [4], [5], [6]. For the education data, the missing data can be caused by the unavailability of internet connections in remote areas so that the school operators are unable to key in the necessary data, number and capacity of human resources, or human error. Three types of missing data, namely Missing Completely At Random (MCAR), Missing At Random (MAR), and Not Missing At Random (NMAR), [7], [8], [9]. MCAR is missing data that occurs randomly and has nothing to do with observed or unobserved responses. Meanwhile, MAR is missing data that occurs because it has something to do with observed responses but is not related to unobserved responses, while MNAR is the loss of observed data on a variable related to observed and unobserved responses. MNAR is categorized as the worst performer of missing data because it can cause severe bias, [10], [11], [12].

For certain data analyses, if the number of observations is large enough and the missing data happens randomly, it can be handled by deleting data on that particular row, [13]. However, in the education data, solving the missing education data by deleting the individual school data should be avoided since it will eliminate the school's identity and the policy maker will not be able to assess the performance of that particular school. The better and more practical solution is to complete the missing information of the corresponding indicator by requesting corresponding schools or using a statistical approach to estimate the missing data based on existing data.

There are many statistical methods to handle the missing data; one is the imputation method, which predicts missing values based on the other variables, [14], [15]. Two types of imputation methods are available; based on statistical models and machine learning, [16], [17], [18]. Two effective and efficient imputation methods using statistical models are statistical models for the imputation are Panel Data Regression and Multiple Imputation by Chained Equations (MICE). [19] conducted research using various imputation methods to handle missing data on simulated data and real data on breast cancer patients, including the multiple regression and MICE methods, [20]. MICE was also applied in the research conducted by [21] The study shows that MICE provides good estimation results and can predict missing data flexibly according to the distribution of actual data. Based on the background, this study aims to compare the application of the Multiple Imputation by Chained Equations (MICE) method and Panel Data Regression to estimate missing data in the case of education data in Indonesia.

METHODS

Imputation Approaches Using Statistical Models

- Panel Data Regression
  The panel data regression model can be expressed in the following equation, (Brüderl & Ludwig, 2015; Hsiao, 2022)

$$y_{it} = \beta_0 + \sum_{k=1}^{P} \beta_k x_{ikt} + u_{it} \quad (1)$$

$$u_{it} = \mu_i + \lambda_t + v_{it} \quad (2)$$

$$i = 1,2,\dots,N;\ t = 1,2,\dots,T;\ k = 1,2,\dots,P$$

  where y_{it} is response of the *i*th individual on the *t*th time, β_0 is intercept, β_k is regression coefficient of the *k*th-explanatory variable, x_{ikt} is the *k*th-explanatory variable for the *i*th individual at the *t*th time, u_{it} is component error on model, μ_i is individual-specific effect, λ_t is time-specific effect and v_{it} is error for the *i*th-individual at the *t*th-time. Three forms of the panel data regression model are

common, fixed, and random effect models, [2].

- Common Effect Models (CEM)
  The model assumes that the regression coefficients for each variable are the same in that no individual or time-specific effect is involved in the model. The common models formulated as follows, [1]:

$$y_{it} = \beta_0 + \sum_{k=1}^{P} \beta_k x_{ikt} + u_{it} \tag{3}$$

$$i = 1,2,\dots,N;\ t = 1,2,\dots,T;\ k = 1,2,\dots,P$$

  where $u_{it} \sim iid(0, \sigma_u^2)$. Regression parameters in the common model are estimated using the ordinary least squares (OLS) method which minimizes the sum of the squared error estimates.

- Fixed Effect Model (FEM)
  In the fixed effect model, the unobserved specific effect is correlated with explanatory variables, [22], and the specific effect is included in the regression parameters to be estimated. The fixed effect model, in general, can be expressed in the following equation, [23].

$$y_{it} = \beta_0 + \mu_i + \lambda_t + \sum_{k=1}^{P} \beta_k x_{ikt} + u_{it} \tag{4}$$

$$i = 1,2,\dots,N;\ t = 1,2,\dots,T;\ k = 1,2,\dots,P$$

  where $\mu_i$and λ_t are separated from the error component of the u_{it}. There are two types of specific effects on the fixed effect model, namely one-way and two-way effects, [1]. In the one-way effect, there will be either one of the individual or time-specific effects in the model, while in the two-way effect, both individual and time-specific effects must be present. In the fixed effect model with individual-specific effects, the model intercept differs between individuals, while in the fixed effects model with time-specific effects, the model intercept differs between time. For fixed effect models with individual and time-specific effects, the model intercept differs between individuals and between time.

- Random Effects Model (REM)
  A random effect model is a model in which unobserved specific effects are assumed to be uncorrelated with explanatory variables, [23]. The random effect model is written as follows:

$$y_{it} = \beta_0 + \sum_{k=1}^{P} \beta_k x_{ikt} + u_{it} \tag{5}$$

$$u_{it} = \mu_i + \lambda_t + v_{it} \tag{6}$$

$$i = 1,2,\dots,N;\ t = 1,2,\dots,T;\ k = 1,2,\dots,P$$

  where $\mu_i \sim iid(0, \sigma_\mu^2)$, $\lambda_t \sim iid(0, \sigma_\lambda^2)$, $v_{it} \sim iid(0, \sigma_v^2)$, μ_i and λ_t are not correlated with v_{it}, while μ_i and λ_t are not correlated with x_{ikt}. In this model, Individual specific effects μ_i and λ_t are included in the error component, so u_{it} will be correlated with each other. As the consequences:

$$Corr(u_{it}, u_{js}) = \sigma_\mu^2/(\sigma_\mu^2 + \sigma_\lambda^2 + \sigma_v^2); i = j, t \neq s \quad (7)$$
$$= \sigma_\lambda^2/(\sigma_\mu^2 + \sigma_\lambda^2 + \sigma_v^2); i \neq j, t = s$$
$$= 1; i = j, t = s$$
$$= 0; i \neq j, t \neq s$$

with $\sigma_\mu^2 = Var(\mu_i)$, $\sigma_\lambda^2 = Var(\lambda_t)$, and $\sigma_v^2 = Var(v_i)$. Estimation of regression parameters with the OLS method cannot be used in the random effect model because the error terms are always correlated. Therefore, the generalized least squares estimation method will be used for the parameter estimation [24]. Chouw test can be used to choose the better model among common and fixed effect models, [23]. Meanwhile, the Hausman test can be implemented to determine the better model between the random and the fixed model, [1]. But the Lagrange Multiplier test can only be used for comparing the common and random effect models, [25].

Multivariate Imputation by Chained Equations (MICE)

The MICE method is also known as a fully conditional specification or sequential regression multiple imputations, which is a method that uses the chain equation approach and is flexible to handle variables with various types of data [26]. MICE can take variables with polychotomous logistic regression data types and proportional odds. It is suitable for MAR-type missing data. Multiple imputations are categorized into three stages: data imputation, data analysis, and pooling, [13]. The data imputation stage is the process of imputing *m* times on data sets containing missing data with several values to produce *m* complete data sets. The data analysis stage is performed for each *m* complete data set resulting from the imputation stage to produce parameter estimators. The final stage is pooling in which the *m* set of parameters obtained at the analysis stage will be calculated using Rubin's rules to produce a final set of parameters [27].

Longitudinal data is a special case of multilevel data so the hierarchical structure of the data will be considered in constructing the imputation model. The MICE can be applied to handle cases of missing data in longitudinal studies. With the MICE, multilevel model-based imputation is conducted iteratively for each variable that contains missing data [13]. For example, denote a measuring vector that contains *i*th individual response variables $\boldsymbol{y}_j$ $(i = 1,2, \dots n_i)$ and class*j* $(j = 1,2, \dots J)$. The mixed effects model is expressed in the following equation:

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{u}_j + \boldsymbol{e}_j \quad (8)$$

with $\boldsymbol{X}_j$ is the design matrix $(n_j \times p)$ of the size of the *j*th class associated with a fixed effect vector $\boldsymbol{\beta}$ $(p \times 1)$, and $\boldsymbol{Z}_j$ is the design matrix $(n_j \times q)$ of the *j*th class associated with a random effect vector $\boldsymbol{u}_j$ $(q \times 1)$. The random effect u_j is assumed to be normally distributed $u_j \sim N(0, \boldsymbol{\Omega})$ with q usually smaller than p. The $\boldsymbol{e}_j$ represents a vector of $(n_j \times 1)$ size that contains the model error where $e_j \sim N(0, \sigma_j^2 \boldsymbol{I}(n_j)\ for\ j = 1,2, \dots J$.

The imputation is carried out in the multilevel model based on the Bayesian approach. Suppose y_j is a variable that contains missing data. It can be stated that: $y_j = [y^{obs}, y^{mis}]$where y^{obs}states the observed data and y^{mis}states missing data. Assuming data is missing according to the MAR mechanism, the parameter distribution can be simulated using the Markov Chain Monte Carlo (MCMC) method with the following steps [28].

1. Take samples β from $p(\beta | y^{obs}, u, \sigma^2)$,

2. Take samples $u_i$from $p(u|\ y^{obs}, \beta, \Omega, \sigma^2)$,
3. Take samples Ω from $p(\Omega|u)$,
4. Take samples σ^2 from $p(\sigma^2|\ y^{obs}, \beta, u)$,
5. Repeat steps 1-4 until you get a convergent result, and
6. Take samples y^{mis} from $p\left(y^{mis}\middle|y^{obs}, \beta, u, \Omega, \sigma^2\right)$

The imputation of $y_i$variables are carried out by taking samples $e_j^* \sim N(0, \sigma^2)$ so that the imputed results are obtained based on the following equation:

$$\boldsymbol{y}_j^* = \boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{u}_j + \boldsymbol{e}_j^* \tag{9}$$

Another approach to estimating the missing data in the imputation method is by estimating the missing data sequentially, called multiple imputation. In multiple imputations, the estimation is conducted by calculating the average of *m* in the complete imputed dataset and considering it the final imputed result. Nevertheless, this is not a reliable, according to [13], since it ignores the variability between each imputed result. In practice, at the imputation stage, *m* complete data sets are produced where each missing data is estimated with a different value. At this stage of the analysis, an analysis was carried out for each *m* complete set of imputed data according to the researcher's preferences. For example, researchers are interested in applying multiple regression to *m* complete imputed data sets which will produce as many as *m* sets of estimated regression parameters [20]. For this purpose, Rubin's rule are employed since it is designed to pool parameter estimates, such as regression coefficients, [29]. When the Rubin's rule are used, it is assumed that the repeated parameter estimates are normally distributed. This cannot be assumed for all statistical test statistics. For a single population parameter of interest, θ, e.g. a regression coefficient, the multiple imputation overall point estimate is the average of the *m* estimates of θ from the imputed datasets, For these test statistics, transformations are first performed before RR can be applied. Then , to calculate the pooled mean of the parameter estimates, the following formula is used:

$$\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m}\widehat{\theta}_i \tag{10}$$

where $\bar{\theta}$ is the estimated regression coefficient of the pooling results and $\widehat{\theta}_i$ is the estimated regression coefficient in the ith complete data set. The standard error in multiple imputations combines two sources of variance: within-imputation variance and between-imputation variance. Within-imputation variance can be calculated using the following formula:

$$V_W = \frac{1}{m}\sum_{i=1}^{m}SE_i^2 \tag{11}$$

where V_W is the within-imputation variance and SE_i^2 is the sample variance from the *i*th complete data set. Between-imputation variance can be calculated using the following equation:

$$V_B = \frac{1}{m-1}\sum_{i=1}^{m}(\widehat{\theta}_i - \bar{\theta})^2 \tag{12}$$

where V_B is the between-imputation variance, $\bar{\theta}$ is the estimated regression coefficient of the pooling results and $\widehat{\theta}_i$is the estimated regression coefficient in the ith complete data

set. Furthermore, the total diversity can be obtained using the following equation:

$$V_T = V_W + V_B + \frac{V_B}{m} \tag{13}$$

If the value of *m* used is close to infinity, then $V_T = V_W + V_B$.

Estimation Accuracy

Several statistics to measure the accuracy of estimation of missing data are Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). MAPE is one of the most commonly used measurements in calculating the accuracy of an estimate, [30], which is written as:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{14}$$

where *n* is the amount of data, y_i is ith actual value and $\hat{y}_i$ is ith estimated value.
A MAPE value of less than 10% indicates that the model used for estimation is very good, MAPE of 10% -20% indicates that the estimation model is good, 20% -50% indicates that it is still feasible to use, while more than 50% indicates poor estimation [31]. Meanwhile, MAE is a measure that describes the diversity of the estimated error values. The small value of diversity illustrates that the model used in the estimation is good. The calculate the MAE, the following equation is used.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{15}$$

where *n* is the amount of data, y_i is ith actual value, $\hat{y}_i$ is ith estimated value.
In addition, RMSE is a measure that describes the standard deviation of the estimated value error. The smaller the value, the better the resulting estimate. The RMSE value is calculated using the following formula [4]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{16}$$

where *n* is the amount of data, y_i is ith actual value, $\hat{y}_i$ is ith estimated value.

Data

Data from 161 randomly selected high schools and vocational schools in Jakarta province from 2016 to 2020 were used. It accumulates to 805 observations (schools) in total. The number of variables in the study was 16 variables Table 1.

Table 1. Variables used in research

Code	Variables	Scale
Type	School type (High school/vocational school)	Nominal
Status	Status (State/Private)	Nominal
Year	School year	Nominal
X1	Average national exam scores	Intervals
X2	Percentage of graduates	Ratio
X3	Percentage of students who drop out	Ratio
X4	Student-to-class ratio	Ratio
X5	Teacher to student ratio	Ratio
X6	Percentage of teachers who have an educator certificate	Ratio
X7	Percentage of teachers who hold a bachelor's degree	Ratio

X8	The ratio of administrative personnel to the number of classes	Ratio
X9	The ratio of the number of classrooms to the number of classes	Ratio
X10	The ratio of the number of computers to the number of students	Ratio
X11	The ratio of the number of students to the number of toilets	Ratio
X12	Laboratory availability	Ratio
X13	Support space availability	Ratio

The Analysis Stages

R studio version 4.2.1 was used for the data analysis with the following steps:

1. Estimation of missing data using panel data regression
   a. Generating missing data on the variables X5, X6, and X7 according to the Missing at Random (MAR) mechanism.
   b. Applying the imputation method using panel data regression for the X5, X6, and X7 variables.
   c. Formulating a common, fixed, and random effect model.
   d. Choosing the better model by conducting the Chow, Hausman, and Lagrange Multiplier tests.
   e. Estimating the missing data using the best panel regression model.
2. Estimation with MICE
   a. Calculating the intraclass correlation coefficient (ICC).
   b. Formulating 5 complete imputed data sets (m=5) with 50 iterations.
   c. Performing analysis on each imputed complete data set.
   d. Pooling the imputation model parameters.
   e. Estimating missing data with MICE.
   f. Comparing panel Regression and MICE results.

RESULTS AND DISCUSSION

The simulation generates 10% missing data on the variables *X5*, X6, and X7 of 805 total observations based on MAR. Eighty missing data were spread across 56 schools, with seven of observations of missing data at X5, nine at X6, and 64 at variable X7.

Implementation of the Imputation Method with Panel Data Regression

Parameter estimation of the CEM model, FEM model, and REM model was carried out for each response variable X5, X6, and X7. The best model was then selected for each response variable using the Chow, Hausman, and Lagrange multiplier tests Table 2.

Table 2. The best model test results

Test	X5	X6	X7
Chow	F = 2.921 (p=0.00)	F = 18.167 (p=0.00)	χ^2_K= 573.48 (p=0.00)
Hausman	χ^2_K= 13.065 (p=0.289)	χ^2_K= 210.67 (p=0.00)	χ^2_K= 21.423 (p=0.00)
LM	χ^2_K= 70.933 (p=0.00)	χ^2_K= 458.94 (p=0.00)	χ^2_K= 114.53 (p=0.00)

The Hausman test results indicate that the best model for the response variable X5 is a random effect model, while based on the Chow and the Hausman test show that X6 and X7 are fixed effect models. Furthermore, the LM test was carried out to see whether there was an individual effect, a time effect, or an individual and time effect on X6 and X7 Table 3.

Table 3. The results of the effect test on the fixed effect models X6 and X7

Effect	X6	X7
Two-way direction	χ^2_K= 573.48 (p=0.00)	χ^2_K= 170.34 (p=0.00)
Individual	χ^2_K= 21.423 (p=0.00)	χ^2_K= 13.016 (p=0.00)
Time	χ^2_K= 114.53 (p=0.00)	χ^2_K= 0.916 (p=0.339)

The variable X6 has individual-specific and time-specific effects, while there are individual-specific effects in X7 (Table 3). Therefore, three final models are obtained to estimate missing data on the variables X5, X6, and X7, namely:

$$\hat{X}_{5it} = 0{,}257 - 0{,}010\text{Schooltype2} - 0{,}025Status2 - 7{,}91 \times 10^{-5}X_1 - 0{,}001X_2 - 0{,}005X_4 - 3{,}14 \times 10^{-4}X_6 - 3{,}40 \times 10^{-5}X_7 + 0{,}020X_8 + 0{,}026X_9 + 5{,}94 \times 10^{-4}X_{10} - 1{,}04 \times 10^{-5}X_{11} + 0{,}005X_{12} - 0{,}008X_{13} \quad (17)$$

$$\hat{X}_{6it} = \mu_i + \lambda_t - 0{,}023X_1 + 0{,}172X_2 + 0{,}162X_4 - 9{,}966X_5 + 0{,}187X_7 - 3{,}012X_8 - 0{,}272X_9 + 6{,}908X_{10} - 0{,}007X_{11} + 0{,}318X_{12} + 1{,}248X_{13} \quad (18)$$

$$\hat{X}_{7it} = \mu_i + 0{,}021X_1 - 0{,}011X_2 + 0{,}038X_4 - 10{,}277X_5 + 0{,}045X_6 - 1{,}786X_8 + 0{,}761X_9 + 5{,}789X_{10} + 0{,}003X_{11} + 1{,}298X_{12} + 1{,}231X_{13} \quad (19)$$

Application of the Imputation Method with MICE

The ICC value determines whether or not to use a hierarchical structure in the imputation process. Referring to [32], if the ICC values are greater than 0.39, they are classified as high. It is concluded that the hierarchical structure is considered when building the regression model for these three variables. This study found that the ICC values of the variables X5, X6, and X7 are high Table 4.

Table 3. ICC values of response variables

Variable	X5	X6	X7
ICC	0.677	0.768	0.403

The next step is to form the five complete imputed data sets and iterate 50 times for each complete data set formed. When imputting the variable X5, for example, the missing data in the X6 and X7 are filled with a randomly generated value. Table 5 shows an example of the estimated value of the simulation results from three observations containing missing data at X5, X6, and X7.

Table 4. Examples of actual values and imputed values

Variable	Actual Data	m^{th} Imputed Values				
		m1	***m2***	***m3***	***m4***	***m5***
X5	0.064	0.062	0.056	0.060	0.065	0.050
X6	80	71,478	60,486	100	72,670	72,215
X7	100	99,465	98,908	97,844	98,864	100

Based on these five complete data sets, mixed effect model estimation was carried out to obtain five regression parameter estimators for variables X4, X5, and X6, then used to calculate pooled parameter estimators Tables 6, 7, and 8.

Table 5. Results of analysis and model parameter pooling for the response variable X5

Variable	Regression coefficient					
	m1	**m2**	**m3**	**m4**	**m5**	**Pooled**
Intercept	0.228	0.226	0.229	0.230	0.231	0.229
Schooltype2	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008
Status2	-0.026	-0.026	-0.027	-0.027	-0.027	-0.027
X1	-3,801×10-5	-4,464×10-5	-4.307×10-5	-4.563×10-5	-4.315×10-5	-4.29×10-5
X2	-9.23×10-4	-9.15×10-4	-9.21×10-4	-9.18×10-4	-9.22×10-4	-9.20×10-4
X4	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004
X6	-3.50×10-4	-3.41×10-4	-3.53×10-4	-3.53×10-4	-3.57×10-4	-3.51×10-4
X7	7.3×10-5	9,339×10-5	7.33×10-5	5.603×10-5	5.035×10-5	6,921×10-5
X8	0.012	0.012	0.012	0.012	0.012	0.012
X9	0.027	0.027	0.027	0.027	0.027	0.027
X10	0.020	0.021	0.020	0.020	0.020	0.020
X11	-1.339×10-5	-1.326×10-5	-1.25×10-5	-1.307×10-5	-1.266×10-5	-1.3×10-5
X12	0.018	0.017	0.018	0.017	0.018	0.018
X13	-0.007	-0.007	-0.007	-0.007	-0.007	-0.007

Table 7. Results of analysis and model parameter pooling for the response variable X6

Variable	Regression coefficient					
	m1	**m2**	**m3**	**m4**	**m5**	**Pooled**
Intercept	31,745	30,838	37,425	36,172	34,488	34,134
Schooltype2	-3,623	-3,544	-3,608	-3,524	-3,494	-3,559
Status2	-30,936	-31,152	-31,905	-31,430	-31,438	-31,372
X1	-0.058	-0.058	-0.056	-0.058	-0.054	-0.057
X2	0.089	0.097	0.088	0.085	0.095	0.091
X4	0.014	0.015	0.020	0.003	0.004	0.011
X5	-32,333	-30,070	-33,482	-33,054	-34,653	-32,718
X7	0.386	0.400	0.346	0.351	0.365	0.370
X8	-1,659	-2,915	-3,322	-3,201	-2,761	-2,772
X9	1,397	1.032	1.127	1.513	1,296	1,273
X10	0.968	1.136	0.634	0.780	0.868	0.877
X11	-0.002	-0.001	-0.002	-0.001	-0.001	-0.001
X12	6,620	6,531	7,340	6,323	7,443	6,851
X13	3,236	3,253	2,478	3,649	3,080	3,139

Table 8. Results of analysis and model parameter pooling for the response variable X7

Variable	Regression coefficient					
	m1	**m2**	**m3**	**m4**	**m5**	**Pooled**
Intercept	87,695	86,730	87,213	86,457	88,010	87,221
Schooltype2	-0.448	-0.245	-0.314	-0.203	-0.359	-0.314
Status2	-0.236	0.119	-0.128	0.177	-0.078	-0.029
X1	-0.003	-0.005	-0.006	-0.013	-0.018	-0.009
X2	0.007	0.001	-0.008	0.016	0.010	0.005
X4	0.129	0.135	0.150	0.157	0.141	0.142
X5	3,489	3,940	3,823	4,215	3,680	3,829
X6	0.066	0.067	0.055	0.057	0.058	0.060
X8	-1,292	-0.155	-0.128	0.269	-0.286	-0.318
X9	0.788	1,069	0.978	0.822	0.869	0.905
X10	-2,797	-2,495	-2,143	-2,869	-2,954	-2,652
X11	0.004	0.003	0.003	0.002	0.003	0.003
X12	0.905	1,399	1.378	2,293	1,696	1,534
X13	2,603	2,042	2,325	1,760	1,810	2.108

The final model used to predict missing data is as follows:

$$\hat{X}_{5it} = 0{,}229 - 0{,}008Schooltype2 - 0{,}027Status2 - 4{,}29 \times 10^{-5}X_1 - 9{,}20 \times 10^{-4}X_2 - 0{,}004X_4 - 3{,}51 \times 10^{-4}X_6 + 6{,}921 \times 10^{-5}X_7 + \quad (20)$$

$$0{,}012X_8 + 0{,}027X_9 + 0{,}020X_{10} - 1{,}3 \times 10^{-5}X_{11} + 0{,}018X12 - 0{,}007X13 + \mu_i + \lambda_t$$

$$\hat{X}_{6it} = 34{,}134 - 3{,}559Schooltype2 - 31{,}372Status2 - 0{,}057X_1 + 0{,}091X_2 + 0{,}011X_4 - 32{,}718X_5 + 0{,}370X_7 - 2{,}772X_8 + 1{,}273X_9 + 0{,}877X_{10} - 0{,}001X_{11} + 6{,}851X_{12} + 3{,}139X_{13} + \mu_i + \lambda_t \quad (21)$$

$$\hat{X}_{7it} = 87{,}221 - 0{,}314Schooltype2 - 0{,}029Status2 - 0{,}009X_1 + 0{,}005X_2 + 0{,}142X_4 + 3{,}829X_5 + 0{,}060X_6 - 0{,}318X_8 + 0{,}905X_9 - 2{,}652X_{10} + 0{,}003X_{11} + 1{,}534X_{12} + 2{,}108X_{13} + \mu_i + \lambda_t \quad (22)$$

Comparing Imputation Methods

The accuracy of the imputation results of the two methods is evaluated based on the MAPE, MAE, and RMSE values. According to [31], a MAPE value of less than 10% indicates that the model is very good; between 10-20%, the model is good, and 20-50% indicates the model is still feasible to use to predict data. Table 8 shows that the MAPE values of the MICE method is feasible for estimating the missing data at X5, quite feasible for X6, and very good for X7. The smaller the MAE and RMSE values indicate that the imputation method produces estimates closer to the actual data. The MICE imputation method produces a smaller MAE value than the imputation method with panel regression in estimating missing data on variables X5, X6, and X7. In addition, the MICE imputation method also produces a much smaller RMSE value than the panel regression method. Thus, this study shows that the MICE imputation method is better for estimating missing data variables X5, X6, and X7.

Table 6. The accuracy value of the MICE method imputation and Panel Data Regression

Variable	MAPE (%)		MAE		RMSE	
	MICE	Panel regression	MICE	Panel regression	MICE	Panel regression
X5	15,185	23,813	0.011	0.019	0.012	0.025
X6	23,397	45,281	8,526	24,224	9,309	29,254
X7	3,240	93,294	2,890	89,926	5,172	90,151

CONCLUSIONS

Missing data in certain cases can be ignored in data analysis by deleting the units of analysis or variables that contain missing data. For other cases, especially for official data, such as school data, missing data must be completed, one of which is through imputation. Many imputation methods can be applied whose accuracy is affected by the characteristics of the data and the type of missing data.

Panel regression and MICE methods are two methods that can be used to estimate missing data in panel data types. The Chow, Hausman, and Lagrange Multiplier tests show that each variable can have a certain appropriate panel regression model, namely the common, fixed, or random effect models. Then, the missing data estimation by panel data regression must be done to the appropriate model. Based on MAPPE, MAE, and RMSE, panel data regression produces a fairly good estimator for certain variables, namely two out of 3 variables. Estimating missing data using the MICE method produces more accurate results than panel data regression for all variables.

REFERENCES

[1] D. N. Gujarati, *Basic econometrics*. Prentice Hall, 2022.

[2] A. Bell and K. Jones, "Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data," *Political Sci Res Methods*, vol. 3, no. 1, pp. 133–153, 2015.

[3] A. R. Alfarisi, H. Tjandrasa, and I. Arieshanti, "Perbandingan Performa antara Imputasi Metode Konvensional dan Imputasi dengan Algoritma Mutual Nearest Neighbor," *Jurnal Teknik ITS*, vol. 2, no. 1, pp. A73–A76, 2013.

[4] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[5] K. M. Lang and T. D. Little, "Principled missing data treatments," *Prevention science*, vol. 19, no. 3, pp. 284–294, 2018.

[6] A. J. Izenman, *Modern multivariate statistical techniques*, vol. 1. Springer, 2008.

[7] M. W. Heymans and J. W. R. Twisk, "Handling missing data in clinical research," *J Clin Epidemiol*, vol. 151, pp. 185–188, 2022.

[8] R. M. Cook, "Addressing missing data in quantitative counseling research," *Counseling Outcome Research and Evaluation*, vol. 12, no. 1, pp. 43–53, 2021.

[9] J. T. Chi, E. C. Chi, and R. G. Baraniuk, "k-pod: A method for k-means clustering of missing data," *Am Stat*, vol. 70, no. 1, pp. 91–99, 2016.

[10] T. F. Johnson, N. J. B. Isaac, A. Paviolo, and M. González-Suárez, "Handling missing values in trait data," *Global Ecology and Biogeography*, vol. 30, no. 1, pp. 51–62, 2021.

[11] G. T. Waterbury, "Missing data and the Rasch model: The effects of missing data mechanisms on item parameter estimation," *J Appl Meas*, vol. 20, no. 2, pp. 154–166, 2019.

[12] D. Feng, Z. Cong, and M. Silverstein, "Missing data and attrition," in *Longitudinal Data Analysis*, Routledge, 2013, pp. 71–96.

[13] S. Van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *J Stat Softw*, vol. 45, pp. 1–67, 2011.

[14] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*, vol. 793. John Wiley & Sons, 2019.

[15] P. Li, E. A. Stuart, and D. B. Allison, "Multiple imputation: a flexible tool for handling missing data," *JAMA*, vol. 314, no. 18, pp. 1966–1967, 2015.

[16] J. M. Jerez *et al.*, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artif Intell Med*, vol. 50, no. 2, pp. 105–115, 2010.

[17] W.-C. Lin and C.-F. Tsai, "Missing value imputation: a review and analysis of the literature (2006–2017)," *Artif Intell Rev*, vol. 53, pp. 1487–1509, 2020.

[18] W.-C. Lin, C.-F. Tsai, and J. R. Zhong, "Deep learning for missing value imputation of continuous data and the effect of data discretization," *Knowl Based Syst*, vol. 239, p. 108079, 2022.

[19] A. M. Gad and R. H. M. Abdelkhalek, "Imputation methods for longitudinal data: A comparative study," *International Journal of Statistical Distributions and Applications*, vol. 3, no. 4, p. 72, 2017.

[20] C. K. Enders, *Applied missing data analysis*. Guilford Publications, 2022.

[21] H. Romaniuk, G. C. Patton, and J. B. Carlin, "Multiple imputation in a longitudinal cohort study: a case study of sensitivity to imputation methods," *Am J Epidemiol*, vol. 180, no. 9, pp. 920–932, 2014.
[22] J. Brüderl and V. Ludwig, "Fixed-effects panel regression," *The Sage handbook of regression analysis and causal inference*, pp. 327–357, 2015.
[23] C. Hsiao, *Analysis of panel data*, no. 64. Cambridge university press, 2022.
[24] K. Mahmud, A. Mallik, M. F. Imtiaz, and N. Tabassum, "The bank-specific factors affecting the profitability of commercial banks in Bangladesh: A panel data analysis," *International Journal of Managerial Studies and Research*, vol. 4, no. 7, pp. 67–74, 2016.
[25] J. M. Wooldridge, *Introductory econometrics: A modern approach*. Cengage learning, 2015.
[26] V. M. Musau, A. G. Waititu, and A. K. Wanjoya, "Modeling panel data: Comparison of GLS estimation and robust covariance matrix estimation," *American Journal of Theoretical and Applied Statistics*, vol. 4, no. 3, pp. 185–191, 2015.
[27] R. Zulfikar and M. M. STp, "Estimation model and selection method of panel data regression: an overview of common effect, fixed effect, and random effect model," *JEMA: Jurnal Ilmiah Bidang Akuntansi*, pp. 1–10, 2018.
[28] J. N. Wulff and L. E. Jeppesen, "Multiple imputation by chained equations in praxis: guidelines and review," *Electronic Journal of Business Research Methods*, vol. 15, no. 1, pp. 41–56, 2017.
[29] G. Chhabra, V. Vashisht, and J. Ranjan, "A comparison of multiple imputation methods for data with missing values," *Indian J Sci Technol*, vol. 10, no. 19, pp. 1–7, 2017.
[30] S. Van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *J Stat Softw*, vol. 45, pp. 1–67, 2011.
[31] J. R. van Ginkel and P. M. Kroonenberg, "Analysis of variance of multiply imputed data," *Multivariate Behav Res*, vol. 49, no. 1, pp. 78–91, 2014.
[32] C. Chen, J. Twycross, and J. M. Garibaldi, "A new accuracy measure based on bounded relative error for time series forecasting," *PLoS One*, vol. 12, no. 3, p. e0174202, 2017.
[33] J. J. M. Moreno, A. P. Pol, A. S. Abad, and B. C. Blasco, "Using the R-MAPE index as a resistant measure of forecast accuracy," *Psicothema*, vol. 25, no. 4, pp. 500–506, 2013.
[34] J. J. Hox, M. Moerbeek, and R. Van de Schoot, *Multilevel analysis: Techniques and applications*. Routledge, 2017.