Comparison of Multiple Regression and Model Averaging Model-Building Approach for Missing Data with Multiple Imputation

Mohd Asrul Affendi Abdullah
Department of Mathematics and Statistics, Faculty of Applied Science and Technology, University Tun Hussein Onn Malaysia, Campus Pagoh, Muar, Johor, Malaysia
afendi@uthm.edu.my

Lai Jessintha
Department of Mathematics and Statistics, Faculty of Applied Science and Technology, University Tun Hussein Onn Malaysia, Campus Pagoh, Muar, Johor, Malaysia
jessinthalai@gmail.com

Gopal Pillay Khuneswari
Department of Mathematics and Statistics, Faculty of Applied Science and Technology, University Tun Hussein Onn Malaysia, Campus Pagoh, Muar, Johor, Malaysia
khuneswari@uthm.edu.my

Siti Afiqah Muhamad Jamil
Faculty of Computer and Mathematical Sciences, University Technology Mara, Shah Alam, Selangor, Malaysia
afendi@uthm.edu.my

Oyebayo Ridwan Olaniran
Department of Statistics, Faculty of Physical Sciences, University of Ilorin, Ilorin, Kwara State, Nigeria
olaniran.or@unilorin.edu.ng (corresponding author)

Received: 3 September 2024 | Revised: 5 October 2024 and 14 October 2024 | Accepted: 16 October 2024

 DOI: https://doi.org/10.48084/etasr.8909

ABSTRACT

Model construction is of significant importance for the extraction of information from datasets and the prediction of responses based on predictor variables. The objective of this study is to compare the Multiple Regression (MR) and model averaging approaches in the context of missing data and to validate the effectiveness of the Multiple Imputation (MI) method used to address missing data issues. A comparison was performed between the results obtained from the multiple-imputed data and those derived from the Complete Case (CC) data, using a diabetes dataset from Hospital Besar Alor Setar. Prior to the application of MI and model building, k-fold cross-validation was employed to partition the dataset, resulting in 90% of the data lacking complete covariates for training and 10% of the data comprising complete covariates for testing. Subsequently, MI was applied to the 90% training dataset. Model M115, derived from the multiple-imputed data, was identified as the optimal model for MR. In the model averaging approach, two models were identified as optimal: Model 1 (without interaction variables) and Model 2 (with interaction variables). The first one, exhibited the lowest values of Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). These results indicate that model averaging, specifically Model 1, is the superior model-building approach for this study, demonstrating improved performance compared to MR and validating the effectiveness of the MI method.

Keywords-statistical modeling; regression analysis; model averaging; missing data; multiple imputation

I. INTRODUCTION

Statistical modeling is a fundamental tool for elucidating the relationships among variables, evaluating hypotheses, and forecasting outcomes. The construction of these models hinges on the model-building process, which frequently employs MR analysis, a widely used method in fields, such as social sciences, economics, and medicine [1]. MR allows researchers to estimate the relationships between a dependent variable and multiple independent variables, thereby facilitating a more profound understanding of correlations, interactions, and potential causal relationships. However, the reliability of these insights is contingent upon the model-building process, which entails the selection of an appropriate model from the multitude of potential alternatives [2]. This phase is frequently challenging, particularly when determining the most pertinent independent variables or addressing complex datasets. The challenge of model selection is typically addressed through the use of forward selection, backward elimination, and stepwise selection, which involve the iterative inclusion or exclusion of variables based on their statistical significance or contribution to model fit [3]. Although these methods are effective to some extent, they are susceptible to overfitting and underfitting, particularly when models are selected based on criteria, such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). These criteria prioritize parsimony in models, balancing the complexity of a model with its goodness of fit [4]. Nevertheless, even with these guidelines in place, the process of selecting an appropriate model remains a subjective one, with the specific goals of the study often exerting a significant influence. More recent research has called into question the robustness of these techniques, particularly in cases where the research objective is unclear or where multiple plausible models could explain the data [5].

A significant drawback of conventional model selection is its exclusive emphasis on identifying a single optimal model, which can result in the neglect of the inherent uncertainty associated with the model-building process. Model averaging represents a promising alternative, as it considers multiple models and averages their estimates, rather than relying on a single model [6]. This approach diminishes the probability of bias resulting from an excessive reliance on a single model and enhances the resilience of statistical inference. Model averaging has been demonstrated to address issues, such as standard error underestimation, thereby facilitating the generation of more accurate parameter estimates and confidence intervals [7]. Notwithstanding its advantages, model averaging is not yet widely adopted, and there are gaps in the literature concerning its application in the presence of missing data. Missing data introduces another layer of complexity into the process of model building, as it is a common issue in real-world datasets. Failure to address this issue can result in biased or inefficient results [8]. Conventional methodologies, such as listwise or pairwise deletion, often result in the exclusion of incomplete cases, which can markedly diminish the sample size and statistical power, particularly in small datasets. Although these methods are simple, they are susceptible to introducing biases if the data are not missing completely at random [9]. In contrast, MI represents a more sophisticated technique, whereby several plausible values are generated for each missing data point, thus creating numerous complete datasets for analysis. This approach has been demonstrated to yield more reliable estimates by maintaining the variability and intrinsic structure of the data [10].

The combination of model averaging with MIs has yet to be extensively explored, despite its potential to offer more robust inferences in the presence of missing data. Previous studies have primarily concentrated on enhancing the precision of MR models through model selection or addressing missing data through imputation techniques. Nevertheless, few studies have conducted a systematic comparison of these methodologies [11]. This gap in the literature is particularly significant, as modern datasets often contain a mix of incomplete and complex variables, necessitating a more comprehensive approach to model building those accounts for both uncertainty in model selection and the complications of missing data to be resolved. The objective of this study is to address this gap by conducting a systematic comparison between MR and model-averaging approaches within the context of MIs for handling missing data. The current paper's contribution is an evaluation of the relative performance of these two model-building strategies, with a focus on their ability to produce reliable and accurate estimates in the presence of incomplete data. By providing a detailed comparison, this study aims to offer practical guidance for researchers and statistical analysts on selecting the most appropriate model-building approach when dealing with missing data, while advancing the theoretical understanding of how these methods perform under different conditions. The findings of the present study have significant implications for fields where the missing data issue is prevalent, including healthcare, social sciences, and economics. The former provide valuable insights into the relative merits and limitations of different model-building strategies.

II. MATERIALS AND METHODS

A. Dataset

The dataset used in this study comprises patient records from the diabetes dataset of Hospital Besar Alor Setar. The primary outcome of interest is plasma glucose concentration, also referred to as blood glucose concentration, which serves as the dependent variable in the conducted analysis. The independent variables are:

- Age: The age of the patient, measured in years.
- Gender: The patient's gender is categorized as male or female.
- Diastolic Blood Pressure: The diastolic blood pressure measurement is recorded in millimetres of mercury (mmHg).
- Systolic Blood Pressure: The systolic blood pressure measurement is also recorded in mmHg.
- Pulse Rate: The patient's pulse rate is measured in beats per minute (bpm).
- Level of Hemoglobin: The hemoglobin concentration in the blood, measured in grams per deciliter (g/dL).

- Presence of Coronary Heart Disease (CHD): A binary variable indicating whether the patient has been diagnosed with coronary heart disease (yes or no).

B. Analysis Approach

The analysis was carried out using two principal approaches:

- A CC analysis was conducted to analyze only those records that contain no missing data for any of the variables of interest. This traditional approach is frequently employed in the presence of missing data, but it can result in biased estimates due to the reduced sample size [12].
- MI generates multiple datasets by incorporating information from other variables in the data set to fill in the gaps in the original data set. In contrast to the inferences produced by simple imputation methods, those derived from MI are more accurate in reflecting the uncertainty associated with missing data. The technique entails imputing each missing value with a vector of imputed values, incorporating random variations through an appropriate model [13, 14].

The MI process mainly involves replacing incomplete data with imputed values on multiple instances (typically 3 to 10 times), which results in the generation of several datasets that are deemed to be complete. Subsequently, the desired statistical analysis is conducted on these datasets using standard complete-data methods. The MI process comprises three principal stages:

- The Multiply Imputed Dataset is generated as unknown missing values are replaced by *J*-independent sets of imputed values drawn from the distribution of the missing data conditional on the observed data.
- The multiply imputed dataset is then subjected to analysis. Once the MIs have been generated, the complete datasets are obtained. Each imputed dataset is then subjected to a separate analysis, with parameters estimated for each. It should be noted that the results will vary depending on the specific imputations that have been applied to replace the missing values.
- The next step is to combine the estimates from the multiply imputed datasets. The *J* estimates are aggregated to form overall estimates in accordance with Rubin's Rules (RR). The combined variance-covariance matrix incorporates both within-imputation and between-imputation variability, thereby providing a more accurate overall forecast.

C. Model Building Approaches

- MR is a regression model comprising more than one independent variable, which is used to describe the behavior of the dependent variable. In other words, it generalizes the simple linear regression model by allowing for more than one term in a mean function, rather than just one intercept and slope. The MR model may include any number of independent variables [15-23]. Accordingly, the linear additive model that describes the relationship between a dependent variable, Y_i, and p independent variables, X_i, is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots\ldots + \beta_p X_{ip} + \varepsilon_i \quad (1)$$

In addition, an example of the regression model for two independent variables (X_1 and X_2) with the first order interaction between X_1 and X_2, which is X_{12}, can be stated as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_{12} + \varepsilon \quad (2)$$

The development of the mathematical model consists of four distinct phases, as shown in Figure 1.

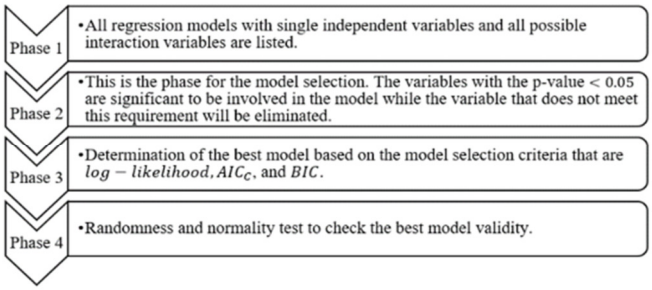

Fig. 1. Four phases of MR for model building approach.

- Model Averaging: Model uncertainty represents a significant challenge that frequently arises during the model-building process. Model selection techniques represent a standard methodology for identifying the optimal model among all potential models. Nevertheless, the selection of a model will typically introduce further uncertainty into the model-building process. This is because the selection of a single model tends to ignore the uncertainty associated with the specification of the selected model. As a result, outcomes may lack precision and the confidence interval may be overly optimistic. Consequently, model averaging represents an alternative methodology to that of model selection. Rather than selecting a single optimal model, the weighted average of the estimates for all potential models is calculated. The application of model averaging will result in more accurate predictions and a reduction in the estimated influence of weaker variables. In model averaging, the "better" model is assigned a higher weight.

The model averaging process can be broken down into five distinct steps. At the initial stage of the process, all potential models will be enumerated. The number of potential models will be equivalent to the number of probable models for the MR procedure. Subsequently, the weight assigned to each potential model will be determined in accordance with the specified criteria for model selection. Once the overall weights have been determined, the averaging estimator or coefficient ($\hat{\beta}_p$) will be calculated. Subsequently, the optimal model will be identified by integrating the estimates pertaining to the model's constituent sets. Lastly, in a manner analogous to the procedure employed in MR, a scatter plot of the residuals for the optimal model will be constructed to ascertain the randomness of the residuals. Furthermore, the skewness and kurtosis values of the optimal model will be calculated to ascertain its normality. In essence, when model averaging is

carried out in conjunction with MI, the model averaging estimator ($\hat{\beta}_p^{(MI)}$) for the linear model can also be defined as:

$$\hat{\beta}_p^{(MI)} = \frac{1}{M}\sum_m^M \hat{\beta}_p^m \quad (3)$$

where M is the total number of models, p is the number of variables in each model, w_j is the weight of each variable j in each model m, and:

$$\hat{\beta}_p^m = \sum_{j=1}^J w_j^m \hat{\beta}_{(p,j)}^m \quad (4)$$

Therefore, the estimated variance ($\widehat{Var}$) is:

$$\widehat{Var}(\hat{\beta}_p^{(MI)}) = \frac{1}{M}\sum_{j=1}^J \left(w_j^m \sqrt{\widehat{Var}\hat{\beta}_{(p,j)}^m + (\hat{\beta}_{(p,j)}^m - \hat{\beta}_p^m)^2} \right)^2 + \frac{M+1}{M(M-1)}\sum_{m=1}^M (\hat{\beta}_p^m - \hat{\beta}_p^{(MI)})^2 \quad (5)$$

D. Evaluation Metrics

Once the optimal model has been identified through the application of both MR and model averaging techniques, it will be subjected to evaluation using established metrics, MSE, RMSE, and MA. These three metrics will be employed to assess the model's performance, with lower values indicating superior performance.

III. RESULTS AND DISCUSSION

Tables I and II present summaries of the variables, providing descriptive statistics for the imputed data. Table II indicates that the mean blood glucose level (Y) is 13.14, while the mean age (X_1) is 59.09. The mean diastolic blood pressure (X_2) is 76.33, and the mean systolic blood pressure (X_3) is 142.32. The mean pulse rate (X_4) is 86.36, and the mean hemoglobin level (X_5) is 11.36. As indicated by the descriptive statistics in Table II, the imputed data closely match the original data, with the mean values for the imputed data being nearly identical to those in Table I. This similarity suggests that the imputed data are valid and suitable for analysis.

TABLE I. SUMMARY OF QUANTITATIVE VARIABLES FOR THE DIABETES DATA

Variable	Mean	Median	Minimum value	Maximum value	Missing value
Y	13.39	12.20	1.20	35.00	114
X_1	57.25	59.00	10.00	94.00	No
X_2	77.38	76.00	38.00	164.00	No
X_3	87.63	87.00	33.00	170.00	No
X_4	11.54	11.40	3.70	21.40	268
X_5	143.00	140.00	16.00	266.00	No

TABLE II. DESCRIPTIVE STATISTICS FOR THE IMPUTED DATASETS

Variable	Mean	Median	Minimum value	Maximum value
Y	13.14	12	1.19	32.43
X_1	59.09	60	27.00	90.00
X_2	76.33	75	38.00	115.00
X_3	86.36	86	44.00	136.00
X_4	11.36	11.30	2.67	18.60
X_6	142.32	140	70.00	216.00

A. MR Model Building Approach for MI Data

The majority of selection criteria indicate that M115 is the optimal model, as evidenced by its AIC and log-likelihood values of 7,757.389 and -3,871.648, respectively. The blood glucose level is 3.7908 mmol/L when the other variables are held constant, as predicted by model M115, and as presented in Table III. However, this is not feasible given that the patients' age, blood pressure, pulse rate, and hemoglobin level cannot be zero. In addition, it can be inferred that an increase of 1 mmHg in blood pressure will result in a reduction of 0.0261 mmol/L in blood glucose levels, and that the presence of coronary heart disease will lead to a further reduction of 2.8806 mmol/L. Furthermore, an increase of 1 bpm in pulse rate will result in an increase of 0.0641 mmol/L in blood glucose level. Additionally, an increase of 1 g/dL in hemoglobin level will lead to an increase of 0.6475 mmol/L in blood glucose level, and an increase of 0.9647 mmol/L in blood glucose level when the patient is female:

$$\hat{Y} = 3.7908 - 0.0261X_2 + 0.0641X_3 + 0.6475X_4 - 0.9647X_{5(Male)} - 2.8806X_{7(With\ CHD)} \quad (6)$$

TABLE III. M115 COEFFICIENTS FOR IMPUTATION DATA

	Model estimator	Std. Error	*P*-value
(Intercept)	3.7908	1.4392	<0.001
X_2	-0.0261	0.01252	<0.001
X_3	0.0641	0.009918	<0.001
X_4	0.6475	0.06788	<0.001
X_5	-0.9647	0.3383	<0.001
X_7	-2.8806	0.33677	<0.001

B. Model Averaging Model Building Approach for MI Data

Model 1 is a statistical model developed using the model averaging approach, which does not take into account variable interactions. According to Model 1, as evidenced in Table IV, the blood glucose level is 3.0982 mmol/L when all other variables are held constant. An one-unit increase in diastolic blood pressure is associated with a decrease of 0.0293 mmol/L in blood glucose level. Furthermore, for each additional year of age, the blood glucose level increases by 0.0002 mmol/L. Similarly, for each 1 bpm increase in pulse rate, the blood glucose level increases by 0.0648 mmol/L:

$$\hat{Y} = 3.0982 + 0.0002X_1 - 0.0293X_2 + 0.0648X_3 + 0.6499X_4 - 0.8996X_{5(Male)} + 0.0057X_6 - 2.8863X_{7(With\ CHD)} \quad (7)$$

TABLE IV. THE MODEL-AVERAGED COEFFICIENT FOR MODEL 1

	Model estimator	Std. Error	*P*-value
(Intercept)	3.0982	1.6875	0.0666
X_1	0.0002	0.0079	0.9816
X_2	-0.0293	0.0209	0.1628
X_3	0.0648	0.0100	<0.001
X_4	0.6499	0.0688	<0.001
X_5	-0.8996	0.3918	0.0218
X_6	0.0057	0.0089	0.5279
X_7	-2.8863	0.3713	<0.001

Model 2 is derived through the application of the model averaging approach, which incorporates variable interactions. According to Model 2, as presented in Table V, the blood glucose level is -1.759 mmol/L when all other variables are held constant. The model predicts that an increase of 1 g/dL in hemoglobin will result in an increase of 3.0567 mmol/L in blood glucose levels. Furthermore, the blood glucose level will increase by 0.0884 mmol/L for each 1-unit increase in the interaction between pulse rate and hemoglobin level. Conversely, the blood glucose level will exhibit a decrease of 0.1813 mmol/L for each 1-unit increase in the interaction between diastolic blood pressure and hemoglobin:

$$\hat{Y} = -1.759 + 1.1737X_1 - 0.8268X_2 + 0.3690X_3 + 3.0567X_4 - 3.1888X_{5(Male)} + 0.3117X_6 - 2.871X_{7(With\ CHD)} + 0.1224X_{12} - 0.0748X_{13} + 0.0781X_{14} + 0.1009X_{23} - 0.1813X_{24} + 0.0884X_{34} + 0.0019X_{123} + 0.0131X_{124} - 0.0146X_{134} - 0.0127X_{234} + 0.0004X_{1234} \quad (8)$$

TABLE V. MODEL-AVERAGED COEFFICIENT FOR MODEL 2

	Estimate	**Std. Error**	***Pr*(>|z|)**
(Intercept)	-1.759	4.2324	0.9101
X_1	1.1737	0.7534	0.6734
X_2	-0.8268	0.6408	0.7228
X_3	0.3690	0.5205	0.4788
X_4	3.0567	3.5618	0.3911
X_5 (male)	-3.1888	1.4620	0.0293
X_6	0.3117	0.4486	0.4874
X_7 (with CHD)	-2.871	1.3902	<0.001
X_{12}	0.1224	0.1134	0.7693
X_{13}	-0.0748	0.0892	0.4018
X_{14}	0.0781	0.1628	0.8964
X_{23}	0.1009	0.0706	0.6977
X_{24}	-0.1813	0.1449	0.7338
X_{34}	0.0884	0.3668	0.8097
X_{123}	0.0019	0.0107	0.9896
X_{124}	0.0131	0.0242	0.8831
X_{134}	-0.0146	0.0513	0.9383
X_{234}	-0.0127	0.0459	0.9400
X_{1234}	0.0004	0.0039	0.9932

C. CC Data Analysis

CC analysis represents the most prevalent technique for addressing issues pertaining to missing data. This study employed CC analysis as an additional methodology for addressing the issue of missing data. The analysis was implemented by removing all cases with missing values and only including all cases with complete values for all variables. In the case of the diabetes dataset, 1,147 CCs were identified, with 346 incomplete cases being subsequently removed from the original dataset. The same model-building approaches were applied to the CC data: MR and model averaging. A comparison of the MSE, RMSE, and MAE indicates that Model A (model averaging without an interaction process) represents the optimal model for CC analysis. Model A exhibited the lowest MSE, RMSE, and MAE.

$$\hat{Y} = 3.2908 + 0.0016X_1 - 0.0224X_2 + 0.0619X_3 + 0.6163X_4 - 0.6736X_{5(Male)} + 0.0068X_6 - 3.1265X_{7(With\ CHD)} \quad (9)$$

In this model (Model A), the blood glucose level is 3.2908 mmol/L when the remaining variables are held constant. Furthermore, an increase of 1 bpm in variable X_3, representing pulse rate, will result in a corresponding increase of 0.0619 mmol/L in blood glucose level. Concurrently, an increase of 0.6163 mmol/L in blood glucose level is observed when variable X_4 (hemoglobin level) is elevated by 1 g/dL. Additionally, blood glucose levels will decrease by 3.1265 mmol/L in the presence of coronary heart disease. The three independent variables with the greatest significance for this model are X_3, X_4, and X_7, as indicated by *p*-values that are less than 0.05.

D. Comparison of Model Building by using MR and Model Averaging

The model building for this study is evaluated through the calculation of the MSE, RMSE, and MAE for each selected model. Table VI demonstrates that the model averaging approach is more effective than MR in building a model, as the evaluation metrics for the former are consistently lower than those for the latter. Therefore, the model averaging approach is deemed the optimal methodology for constructing the model for the diabetes data set. Table VII presents the MSE, RMSE, and MAE values for Model A and Model 1, which are the optimal predictors identified through the model averaging process in both the CC analysis and MI analysis. As illustrated in Table VII, the MSE, RMSE, and MAE values for Model A are higher than those for Model 1. This indicates that MI is an effective approach for addressing missing data. In addition, the systematic imputation technique employed to fill in missing values has been demonstrated to reduce the MSE in comparison to other methods. Moreover, MI preserves the inherent variability of missing values and accounts for their uncertainty, thereby facilitating more precise and reliable statistical inferences.

TABLE VI. EVALUATION METRICS FOR STATISTICAL MODELS (MI DATA)

	MR	**Model 1 (Model averaging without interaction)**	**Model 2 (Model averaging with interaction)**
MSE	32.4832	32.3371	32.8099
RMSE	5.6994	5.6866	5.7280
MAE	4.8674	4.8625	4.8872

TABLE VII. EVALUATION METRICS FOR THE BEST MODEL FROM CC ANALYSIS AND MI ANALYSIS

	Model A (Model Averaging for CC analysis)	**Model 1 (Model averaging for imputed data analysis)**
MSE	32.5482	32.3371
RMSE	5.7051	5.6866
MAE	4.9076	4.8625

IV. FINDINGS

This study compared the efficacy of MR and model averaging approaches for analyzing datasets with missing data, using both CC data and MI data. The results substantiate the efficacy of MI as a robust technique for handling missing data, providing more accurate and less biased estimates in comparison to methods that rely solely on CC analysis. The

findings demonstrate that the model derived from the model averaging process applied to the MI data yielded the lowest MSE, RMSE, and MAE, thereby showcasing its superior performance. In contrast, CC analysis, which excluded data with missing values, proved to be a less reliable method, particularly when the proportion of missing data was high. This further emphasizes the importance of MI in preserving data integrity. A distinctive aspect of this study is its comprehensive examination of model-building techniques in the context of missing data. While previous studies have primarily focused on either model selection techniques or missing data handling separately, this study addresses the limitations of these approaches by incorporating model averaging with MIs. This combination addressed the uncertainty associated with model selection and resolved the common issue of bias that arises in missing data scenarios. By comparing these approaches, the study offers new insights into how model averaging can outperform traditional regression methods in predictive accuracy and model reliability, particularly when applied to datasets with incomplete information.

Moreover, the study's contribution extends beyond mere comparison. It underscores the importance of model averaging in accounting for model uncertainty, a factor that is frequently neglected in conventional MR methodologies. These findings have significant implications for the broader field of statistical modeling, where the choice of model can have a substantial impact on the validity of research findings. Model averaging consistently yielded the lowest values for MSE, RMSE, and MAE, indicating that it provides a more dependable methodology for researchers confronted with missing data. Consequently, it is the preferred approach in instances where model uncertainty and data incompleteness are intricately intertwined. Another noteworthy finding is the discrepancy in performance between CC analysis and MI, which lends further support to the growing consensus that relying on CCs alone can lead to erroneous conclusions. As this study presents, the CC approach becomes increasingly unreliable as the proportion of missing data increases, introducing biases and reducing the overall precision of the model. This contribution adds to the growing body of evidence supporting the use of more advanced techniques, such as MI, which maintain the sample size and data variability, thereby leading to more accurate model estimates. In practical terms, the results of this study provide researchers and practitioners across various disciplines with actionable guidance, particularly in fields, such as social sciences, healthcare, and economics, where missing data is a common occurrence. The demonstrated advantages of combining MIs with model averaging provide a clear recommendation for future research and data analysis practices. This approach enhances the robustness of statistical models and the reliability of conclusions drawn from incomplete datasets, hence ensuring that critical decisions are informed by sound, unbiased evidence.

V. CONCLUSIONS

In conclusion, the novel aspect of this study is its integrated approach to addressing model uncertainty and missing data, which offers a more comprehensive solution than traditional methods. The research demonstrates the superiority of model averaging with Multiple Imputations (MIs), advancing the best practices in statistical modeling with far-reaching implications for the construction and validation of models in incomplete data.

ACKNOWLEDGEMENT

The authors would like to thank the Ministry of Higher Education and the University Tun Hussien Onn Malaysia for supporting this research project.

REFERENCES

[1] B. Hans and E. A. Prasetio, "Applying Multiple Linear Regression Method to Measure the Impact of Human Capital, Social Media, Business Sector and Founder Gender on Advanced-Stage Startup Funding in Indonesia," *Applied Quantitative Analysis*, vol. 3, no. 2, pp. 86–100, Dec. 2023, https://doi.org/10.31098/quant.2039.

[2] S. S. Henley, R. M. Golden, and T. M. Kashner, "Statistical modeling methods: challenges and strategies," *Biostatistics & Epidemiology*, vol. 4, no. 1, pp. 105–139, Jan. 2020, https://doi.org/10.1080/24709360.2019.1618653.

[3] S. Buscemi and A. Plaia, "Model selection in linear mixed-effect models," *AStA Advances in Statistical Analysis*, vol. 104, no. 4, pp. 529–575, Dec. 2020, https://doi.org/10.1007/s10182-019-00359-z.

[4] B. Langenberg, J. L. Helm, and A. Mayer, "Bayesian Analysis of Multi-Factorial Experimental Designs Using SEM," *Multivariate Behavioral Research*, vol. 59, no. 4, pp. 716–737, Jul. 2024, https://doi.org/10.1080/00273171.2024.2315557.

[5] T. Köhler, M. Rumyantseva, and C. Welch, "Qualitative Restudies: Research Designs for Retheorizing," *Organizational Research Methods*, Dec. 2023, Art. no. 10944281231216323, https://doi.org/10.1177/10944281231216323.

[6] C. F. Falk and M. Muthukrishna, "Parsimony in model selection: Tools for assessing fit propensity," *Psychological Methods*, vol. 28, no. 1, pp. 123–136, Feb. 2023, https://doi.org/10.1037/met0000422.

[7] K. Barigou, P.-O. Goffard, S. Loisel, and Y. Salhi, "Bayesian model averaging for mortality forecasting using leave-future-out validation," *International Journal of Forecasting*, vol. 39, no. 2, pp. 674–690, Apr. 2023, https://doi.org/10.1016/j.ijforecast.2022.01.011.

[8] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed. John Wiley & Sons, 2019.

[9] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ, USA: Wiley-Interscience, 2004.

[10] C. K. Enders, *Applied Missing Data Analysis*, 1st ed. New York, USA: The Guilford Press, 2010.

[11] A. Remiro-Azócar, A. Heath, and G. Baio, "Model-based standardization using multiple imputation," *BMC Medical Research Methodology*, vol. 24, no. 1, Feb. 2024, Art. no. 32, https://doi.org/10.1186/s12874-024-02157-x.

[12] T. Z. Keith, *Multiple Regression and Beyond: An Introduction to Multiple Regression and Structural Equation Modeling*, 3rd ed. New York, USA: Routledge, 2019.

[13] D. B. Rubin, "Multiple imputation," in *Flexible imputation of missing data*, 2nd ed., Chapman and Hall/CRC, pp. 29–62.

[14] J. Carpenter and M. Kenward, *Multiple Imputation and its Application*, 1st ed. Chichester, West Sussex, UK: Wiley, 2013.

[15] S. S. Avtar, G. P. Khuneswari, A. A. Abdullah, J. H. McColl, C. Wright, and G. M. S. Team, "Comparison between EM Algorithm and Multiple Imputation on Predicting Children's Weight at School Entry," *Journal of Physics: Conference Series*, vol. 1366, no. 1, Nov. 2019, Art. no. 012124, https://doi.org/10.1088/1742-6596/1366/1/012124.

[16] B. Y. Gravesteijn, C. A. Sewalt, E. Venema, D. Nieboer, E. W. Steyerberg, and the CENTER-TBI Collaborators, "Missing Data in Prediction Research: A Five-Step Approach for Multiple Imputation, Illustrated in the CENTER-TBI Study," *Journal of Neurotrauma*, vol. 38, no. 13, pp. 1842–1857, Jul. 2021, https://doi.org/10.1089/neu.2020.7218.

[17] W. Kim, W. Cho, J. Choi, J. Kim, C. Park, and J. Choo, "A Comparison of the Effects of Data Imputation Methods on Model Performance," in *2019 21st International Conference on Advanced Communication Technology (ICACT)*, PyeongChang, South Korea, Feb. 2019, pp. 592–599, https://doi.org/10.23919/ICACT.2019.8702000.

[18] J. N. Wulff and L. Ejlskov, "Multiple Imputation by Chained Equations in Praxis: Guidelines and Review," *Electronic Journal of Business Research Methods*, vol. 15, no. 1, p. 41-56, Apr. 2017.

[19] B. J. A. Mertens, E. Banzato, and L. C. de Wreede, "Construction and assessment of prediction rules for binary outcome in the presence of missing predictor data using multiple imputation and cross-validation: Methodological approach and data-based evaluation," *Biometrical Journal*, vol. 62, no. 3, pp. 724–741, May 2020, https://doi.org/10.1002/bimj.201800289.

[20] C. D. Nguyen, J. B. Carlin, and K. J. Lee, "Practical strategies for handling breakdown of multiple imputation procedures," *Emerging Themes in Epidemiology*, vol. 18, Apr. 2021, Art. no. 5, https://doi.org/10.1186/s12982-021-00095-3.

[21] L. T. P. Thao and R. Geskus, "A comparison of model selection methods for prediction in the presence of multiply imputed data," *Biometrical Journal*, vol. 61, no. 2, pp. 343–356, Mar. 2019, https://doi.org/10.1002/bimj.201700232.

[22] O. R. Olaniran and M. A. A. Abdullah, "Bayesian weighted random forest for classification of high-dimensional genomics data," *Kuwait Journal of Science*, vol. 50, no. 4, pp. 477–484, Oct. 2023, https://doi.org/10.1016/j.kjs.2023.06.008.

[23] O. R. Olaniran and A. R. R. Alzahrani, "On the Oracle Properties of Bayesian Random Forest for Sparse High-Dimensional Gaussian Regression," *Mathematics*, vol. 11, no. 24, Jan. 2023, Art. no. 4957, https://doi.org/10.3390/math11244957.