Full Information Multiple Imputation for Linear Regression Model with Missing Response Variable

Limin Song, Guangbao Guo

Abstract—**Linear regression models are commonly used to determine the quantitative relationships between variables and utilize the resulting regression equations to make predictions. This paper proposes a fully informative multiple imputation method based on a linear regression model with a missing response variable, utilizing all observable data to obtain estimates of the regression coefficients and thereby the predicted values of the missing response variable. This not only provides a good explanation of the relationship between the response variable and their respective variables, but also effectively enhances the imputation accuracy of the response variable. The stability and sensitivity of the fiMI method are evaluated through a simulation study. Subsequently, the proposed method is applied to two real data sets, the admission prediction data set and the goalkeeper data set, and is discussed and analyzed.**

Index Terms—**linear regression models, missing response variables, full information, multiple imputation.**

I. Introduction

We consider the following linear regression model

$$Y = X\beta + \varepsilon, \tag{1}$$

where $X = (X_{ij}) \in R^{n\times p}$ is the independent variable, $X_{i.} = (X_{i1}, X_{i2}, \cdots, X_{ip})$ represents the $i-th$ row of matrix $X(i = 1, \cdots, n)$, $X_{.j} = (X_{1j}, X_{2j}, \cdots, X_{nj})^\top$ represents the $j-th$ row of matrix $X(j = 1, \cdots, p)$, $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^\top \in R^{p\times 1}$ is a vector of unknown parameters, $Y = (Y_1, Y_2, \cdots Y_n)^\top \in R^{n\times 1}$ is the response variable, and $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots \varepsilon_n)^\top \in R^{n\times 1}$ is the residual vector. $\varepsilon_i \sim N(0, \sigma^2 I_n)$ and independent of each other.

Suppose there are imperfectly independent and identically distributed samples $\{(X_{i.}, Y_i, \delta_i), 1 \le i \le n\}$,where $\{X_{i.}, 1 \le i \le n\}$ is fully observable, $\{Y_i, 1 \le i \le n\}$ is missing, and δ_i is the variable indicating that Y_i is missing, i.e.

$$\delta_i = \begin{cases} 0, \text{if} \quad Y_i \quad \text{is} \quad \text{missing;} \\ 1, \text{if} \quad Y_i \quad \text{is} \quad \text{not} \quad \text{missing.} \end{cases}$$

Assume that Y satisfies the MAR mechanism, i.e.

$$P(\delta_i = 1|X_{i.}, Y_i) = P(\delta_i = 1|X_{i.}, Y_i) = P(X_{i.}),$$

i.e. under a given $X_{i.}$, Y_i is conditionally independent of δ_i.

The number of cells in the response variable Y with no missing data and the number of cells with missing data to be denoted $n_{\text{OB}} = \sum_{i=1}^n \delta_i$ and $n_{\text{NA}} = n - n_{\text{OB}}$, respectively. Define the observable and missing values in the response variable Y to be denoted Y_{obs} and Y_{mis}, respectively, the parts of the matrix X corresponding to Y_{obs} and Y_{mis} to be denoted X_{obs} and X_{mis}, respectively.

For addressing the imputation problem of missing response variables in linear regression models, the most common methods are mean imputation and regression imputation, but these approaches also have some disadvantages. For instance, mean imputation can reduce the correlation between variables, while regression imputation can artificially increase this correlation. Wang et al. [1] (2009) used the expectation and maximization (EM) method to calculate the asymptotic variances and standard errors of the maximum likelihood estimator (MLE) for linear models with missing data for the missing response variable. However, the standard deviation can only be calculated after the operations have converged and cannot be obtained directly. Liu (2012) proposed a new expectation recursive least squares (ERLS) method based on the EM algorithm for linear regression models. Avoiding the difficulty of finding the inverse of the correlation matrix of high-dimensional data. However, the calculation of regression coefficients requires several iterations, which increases the computational time.

The method for dealing with missing data has undergone two main methods: single imputation and multiple imputation. The emergence of multiple imputation methods has addressed the shortcomings of single imputation. Rubin [4] (1987) proposed a multiple imputation procedure that involves replacing each missing data point with a range of potential data sets (thus also reflecting the uncertainty associated with the imputed values); subsequent to this, analyzing these multiple imputed data sets using standard procedures applicable to complete data sets; and ultimately generalizing and consolidating the findings from these analyses. Buuren et al. [2] (2011) used the R package mice to impute incomplete multivariate data using chained equations, providing a practical step-by-step approach to addressing the issue of missing data in applications. The mice package is commonly used to impute missing response variables under linear regression models, with the most commonly used methods being predictive mean matching multiple imputation (PMMMI) method, bayesian multiple imputation (BayesMI) method, and bootstrap multiple imputation (bootstrapMI) method. Rubin [6] (1999) and Schafer [7] (1997) have conducted a series of studies on Bayesian multiple imputation methods, where the imputation accuracy is strongly influenced by the missing data mechanism. Little [8] (1988), Morris et al. [9] (2015), and Buuren [10] (2018) further discussed the predictive mean multiple imputation methods and found that the missing data mechanism has a small impact on the imputation accuracy. Chang et al. [5] (2020) studied the problem of missing data for independent variables

Manuscript received July 1, 2023; revised November 10, 2023.

This work was supported by a grant from National Social Science Foundation Project under project ID 23BTJ059, a grant from Natural Science Foundation of Shandong under project ID ZR2020MA022, and a grant from National Statistical Research Program under project ID 2022LY016.

Limin Song is a postgraduate student of Mathematics and Statistics, Shandong University of Technology, Zibo, China. (e-mail: songsongsonglm@163.com).

Guangbao Guo is a professor of Mathematics and Statistics, Shandong University of Technology, Zibo, China (Corresponding author to provide phone:15269366362; e-mail: ggb11111111@163.com).

in a distributed methods environment, and developed an efficient distributed multiple imputation method for horizontally divided incomplete data communication. However, no solution is provided when the response variable is missing.

II. Full information multiple imputation

Multiple imputation (MI) is arguably the most popular method for dealing with missing data. The MI method replaces each missing value with a sample from its posterior predictive distribution. The predictive imputation model is estimated from the observed data and does not use the missing values. The missing values are imputed multiple times in order to account for the uncertainty of the imputation, and each imputed data set is then used to fit an analysis model. The parameter estimate β is combined with the results of these analyses to produce a final estimate from multiple imputed data sets. This method yields estimates that are more robust than those obtained by using a single value to fill in for the missing data.

A straightforward method to analyzing data is to aggregate information from the minimum observable data so that it will impute by analyzing all observable data. We refer to this method as full information (fi) method, and next we will extend it to the full information multiple imputation (fiMI) method. In linear regression models with missing response variable, the general linear regression imputation requires only $X_{\text{obs}}^{\top}X_{\text{obs}}$ and $X_{\text{obs}}^{\top}Y_{\text{obs}}$ to obtain least squares estimates of the regression coefficients, as can be seen from the following equation:

$$\hat{\beta} = (X_{\text{obs}}^{\top}X_{\text{obs}})^{-1}(X_{\text{obs}}^{\top}Y_{\text{obs}}). \tag{2}$$

However, the regression coefficients estimated in equation (2) may suffer from overfitting, leading to inaccurate predictions. To address this issue, we propose to fit a linear regression imputation model using the fi method, which can be interpreted as fitting the imputation model using all observable data. By passing the imputed model parameters to the full observable data set, it is expected to achieve the best computational performance because it fully exploits all available information.

According to (1), it follows that $Y_i \sim N(X_{i.}\beta, \sigma^2)$ with priors

$$\pi(\sigma^2) \propto IG(1/2, 1/2),$$
$$\beta \mid \sigma^2 \sim N(0, \sigma^2\lambda^{-1}I),$$

where IG and N are denoted as inverse gamma and multivariate Gaussian distributions, respectively. The posterior distribution of (σ^2, β) is given by

$$\begin{aligned} &\sigma^2|X_{\text{obs}} \sim IG((n_{\text{OB}}+1)/2, (SSE+1)/2),\\ &\beta|\sigma^2, X_{\text{obs}} \sim N((X_{\text{obs}}^{\top}X_{\text{obs}}+\lambda I)^{-1}X_{\text{obs}}^{\top}Y_{\text{obs}},\\ &\sigma^2(X_{\text{obs}}^{\top}X_{\text{obs}}+\lambda I)^{-1}) \end{aligned} \tag{3}$$

where

$$SSE = \|Y_{\text{obs}} - X_{\text{obs}}\beta^*\|_2^2,$$

the specific representation of β^* will be given later. The fiMI method samples (σ^2, β) from (3), imputes the missing values of the response variable from (1), and fits the analytical linear regression model using the estimated complete data. This process is repeated m times. To avoid extraneous complexity, we assume that $n_{\text{OB}} > p$.

First, we calculate the matrix

$$A = X_{\text{obs}}^{\top}X_{\text{obs}} + \lambda I_{p\times p},$$

where λ is the regularization parameter, which allows a limited solution to the over-fitting problem in (2). The regression weights

$$\beta^* = (A)^{-1}X_{\text{obs}}^{\top}Y_{\text{obs}}$$

are obtained with reference to (2) and the matrix A. Next, Choleskey's decomposition of the positive definite matrix A yields matrix C_A, i.e.

$$A = (C_A^{\top}C_A),$$

where C_A is the upper triangular matrix. We obtain estimates of the regression coefficients as follows:

$$\hat{\beta} = \beta^* + \sigma(C_A)^{-1}g, \tag{4}$$

where $g = (g_1, g_2, \cdots, g_p)^{\top}$ is a $g_i \sim N(0,1)$ and mutually independent $p-$ dimensional variable. At this point

$$\hat{\beta}_{fi} = \hat{\beta},$$

$$\text{Cov}(\hat{\beta}_{fi}) = \frac{1}{p-1}\sum_{i=1}^{p}((\hat{\beta}_{fi})_i - \bar{\hat{\beta}}_{fi}).$$

According to sufficient statistics $\hat{\beta}_{fi}$ and $\text{Cov}(\hat{\beta}_{fi})$ of the normal distribution, samples $\beta_1, \cdots, \beta_M$ are obtained as being independent of each other and obeying $N(\hat{\beta}_{fi}, \text{Cov}(\hat{\beta}_{fi}))$. Send the multiple regression coefficients $\beta_1, \cdots, \beta_M$ to the imputation model and integrate the multiple imputation results using Rubin's rule to obtain $\hat{\beta}$ and $\text{Cov}(\hat{\beta})$. Based on the final obtained $\hat{\beta}$, impute the missing values $\hat{Y}_{\text{mis}} = X_{\text{mis}}\hat{\beta}$ of the response variable and expand to obtain $\hat{Y}$.

III. Numerical Analysis

A. Evaluation indicators

1) Mean square error of $\hat{Y}$

The mean square error (MSE) calculates the difference between the imputed value and the original true value.

$$\text{MSE}(\hat{\text{Y}}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

where Y_i and $\hat{Y}_i$ denote the original true value and the imputed value respectively.

2) Mean absolute error of $\hat{Y}$

The mean absolute error (MAE) is the average of the absolute differences between each predicted value and the corresponding actual value.

$$\text{MAE}(\hat{\text{Y}}) = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i|.$$

When the difference between the predicted and true values is smaller, it means that the imputation is better.

B. Simulation

Firstly, the initial parameters have been fixed at $(n, p, MR) = (1000, 5, 10\%)$, and the values of $\text{MSE}(\hat{\text{Y}})$ and $\text{MAE}(\hat{\text{Y}})$ have been calculated for the fiMI method and the comparison method under missing response variables. According to Table I, when there is a missing response variable in the linear regression model, the fiMI method has the lowest values for MSE and MAE. Overall, for imputation of linear regression models with missing response variables, the fiMI method has the highest imputation accuracy for the parameter combination $(n, p, MR) = (1000, 5, 10\%)$, meaning that the imputed values from the fiMI method are closest to the true values.

TABLE I
MSE AND MAE VALUES OF FIMI METHOD AND COMPARISON METHODS IN SIMULATED DATA

Indicators	fiMI	ERLS	EMRE	PMMMI	BayesMI	bootstrapMI
MSE (10^{-4})	**1.0472**	1.2040	1.224	1.7151	2.3004	2.1565
MAE (10^{-2})	**8.0541**	8.5226	8.6269	1.0401	1.2393	1.1368

Next, the parameters (n, p, MR) are varied to examine the MSE, MAE, and MRE values of the fiMI method and the comparison method under different sample sizes, numbers of variables, and missing ratios for sensitivity and stability analysis.

Case 1. Varying n with fixed (p, MR)

The parameter values are set as $(p, MR) = (5, 10)\%$ and $n = (300, 500, 1000, 1500, 2000)$. The comparison results are shown in Fig. 1 and Fig. 2.

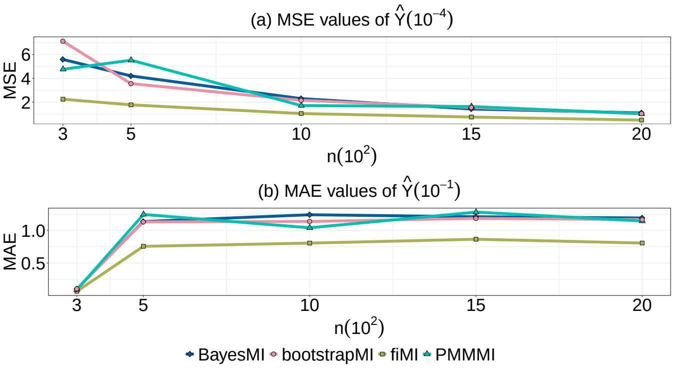

Fig. 1. Results of MSE and MAE values obtained by the fiMI method and multiple comparison methods in simulated data with different n values (case 1)

Upon observing Fig. 1(a) and 2(a), it is found that the MSE value of the fiMI method gradually decreases from 2.2493e-04 to 4.9658e-05 as the sample size n increases for the fixed parameter (p, MR); the MSE values of all other comparison methods are noticeably higher than the fiMI method. Observing Fig. 1(b) and 2(b) reveals that the MAE value of the fiMI method tends to flatten out, indicating that the increase of the sample size n does not have much impact on the MAE value. The fluctuation of the MAE value of the fiMI method is the smallest, and the MAE values of the other methods are higher than the fiMI method.

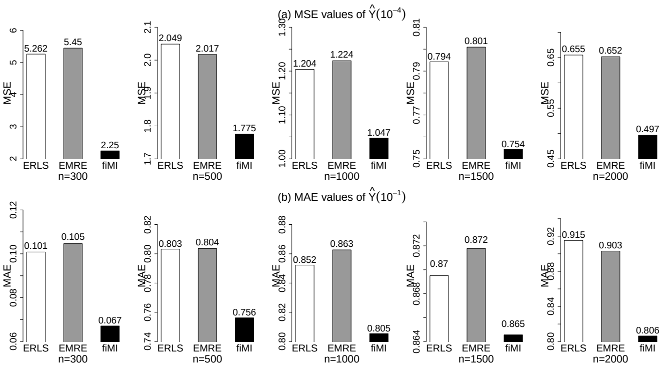

Fig. 2. Results of MSE and MAE values obtained by fiMI, ERLS, and EMRE methods in simulated data with different n values

Case 2. Varying p with fixed (n, MR)

The parameter values are set as $(n, MR) = (1000, 10\%)$ and $p = (3, 5, 10, 15, 20)$. The comparison results are shown in Fig. 3 and Fig. 4.

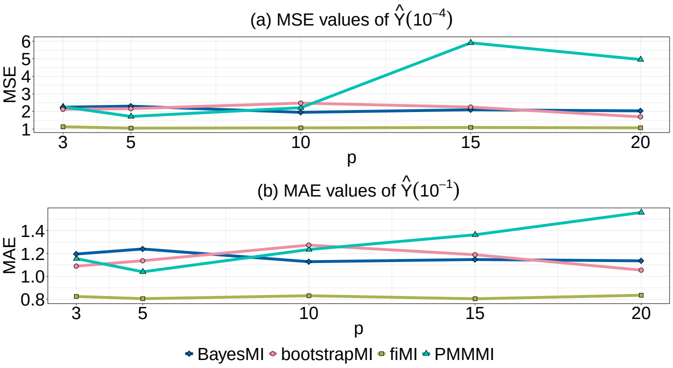

Fig. 3. Results of MSE and MAE values obtained by fiMI method and multiple comparison methods in simulated data with different p values (case 2)

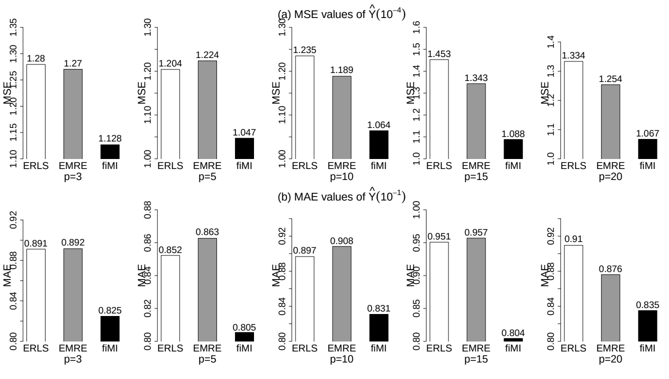

Fig. 4. Results of MSE and MAE values obtained by fiMI, ERLS, and EMRE methods in simulated data with different p values

Upon observing Fig. 3(a) and Fig. 4(a), it is found that the MSE value of the fiMI method fluctuates within the range of 1.047241e-04 to 1.12679e-04 as the number of variables p increases for the fixed parameter, indicating that it is less affected by the number of variables p. The MSE value of the other compared methods is higher than the fiMI method. Observing Fig. 3(b) and 4(b) reveals that the change in the MAE value of the fiMI method is more gradual, indicating that the increase in the number of variables p has less impact on the MAE value. The MAE value of the fiMI method fluctuates within the range of 8.04380e-02∼8.3516e-02, with the smallest range of fluctuation; the MAE values of the other methods are higher than the fiMI method.

Case 3. Varying MR with fixed (n, p)

The parameter values are set as $(n, p) = (1000, 5)$ and $MR = (10\%, 20\%, 30\%, 40\%, 50\%)$. The results are shown in Fig. 5 and Fig. 6.

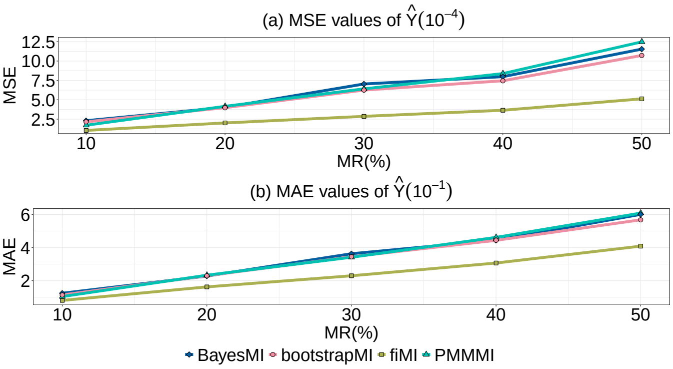

Fig. 5. Results MSE and MAE values obtained by fiMI method and multiple comparison methods in simulated data with different MR values (case 3)

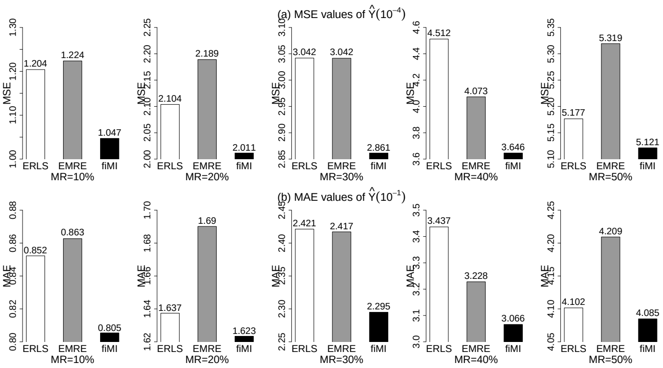

Fig. 6. Results of MSE and MAE values obtained by fiMI, ERLS, and EMRE methods in simulated data with different MR values

Upon observing Fig. 5(a) and Fig. 6(a), it can be seen that the MSE values of both the fiMI method and the other comparative methods show an overall increasing trend with the increase of the MR for the fixed parameter (n, p), with the MSE value of the fiMI method fluctuating within the range of 1.0472e-04 to 5.1215e-04; the other imputation methods have higher MSE values than the fiMI method. Observing Fig. 5(b) and Fig. 6(b) reveals that the MAE values of both the fiMI method and the other comparative methods roughly increase linearly with the increase of the MR, but the MAE value of the fiMI method is the smallest among all the imputation methods, and the fluctuation range is also the smallest.

C. Real Data Analysis

In this section, two real data sets are selected: the admission prediction data set and the goalkeeper data set, and the data set for this empirical study is obtained from a third-party data science community, the Heywhale. The response variable is the chance of admission in the admission prediction data set. Firstly, correlation analysis is done for each variable of the admission prediction data set as shown in Table II:

TABLE II
THE CORRELATION COEFFICIENT AND P-VALUE BETWEEN THE INDEPENDENT VARIABLES AND RESPONSE VARIABLE IN ADMISSION PREDICT DATA SET

Statistical tests	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA
CC	0.803	0.792	0.711	0.676	0.670	0.873
P-value	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16

The correlation and significance test analysis show that these six characteristic variables are all highly correlated with the response variable chance of admission, so the above six characteristic variables are selected as independent variables. The admission prediction data set is suitable for establishing a multiple linear regression model, our regression model is

$$Y_i = \sum_{j=1}^{6} X_{ij}\beta_j + \varepsilon_i, i = 1, 2, \cdots 400.$$

For the admission prediction data set, we set the MR of admission chances to 50% , then impute with the fiMI method and comparison method, and finally compare the imputation methods in terms of imputation accuracy.

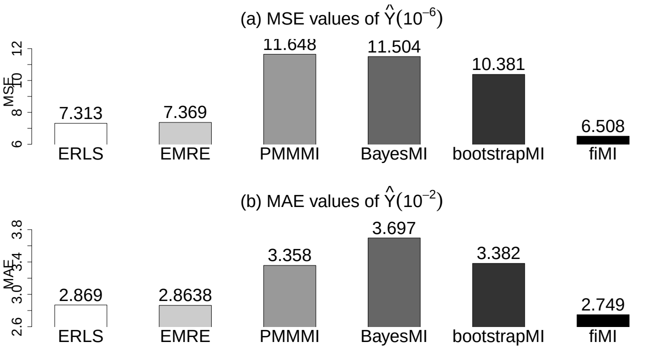

Fig. 7. MSE and MAE values obtained by fiMI method and comparison methods in admission predict data set

It can be seen from Fig. 7 that the fiMI method has the lowest MSE and MAE values for the response variable admission chances of $MR = 50\%$, indicating that the fiMI method has the best imputation effect. Overall, for the admission prediction data set with a large ratio of missing values, the fiMI method has the highest imputation accuracy, followed by the ERLS and EMRE methods.

The second data set for the empirical study is the goalkeeper player data set. Rating is the response variable.

Firstly, we perform correlation analysis on each variable of the goalkeeper data set, and the correlation coefficients and p-values between each characteristic variable and the response variable are calculated as shown in Table III below:

TABLE III
THE CORRELATION COEFFICIENT AND P-VALUE BETWEEN THE INDEPENDENT VARIABLES AND RESPONSE VARIABLE IN GK DATA SET

Statistical tests	Positioning	Diving	Kicking	Handling	Reflexes
CC	0.923319	0.9217224	0.7543833	0.9113288	0.9262662
P-value	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16

The correlation and significance test analysis show that these five characteristic variables are all strongly correlated with rating, so the above five characteristic variables are selected as independent variables. The goalkeeper data set is suitable for multiple linear regression modeling, so $p = 5$, our regression model is as follows:

$$Y_i = \sum_{j=1}^{5} X_{ij}\beta_j + \varepsilon_i, i = 1, 2, \cdots 2003.$$

For the goalkeeper data set, we still consider the case of a large percentage of missing response variables and set the missing ratio of the response variable rating $MR = 50\%$ then impute with the fiMI method and the comparison method, and finally compare the imputation methods in terms of imputation accuracy.

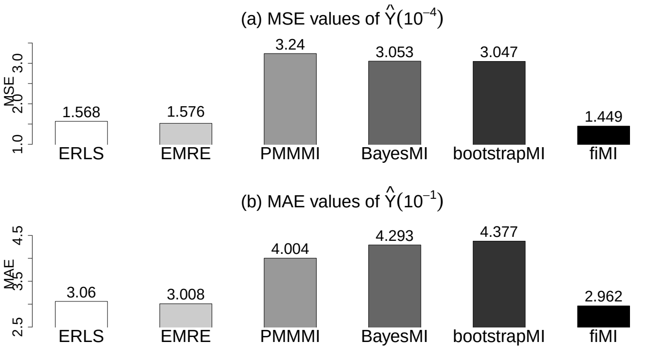

Fig. 8. MSE and MAE values obtained by fiMI method and comparison methods in GK data set

Observation of Fig. 8 reveals that the fiMI method has the lowest MSE and MAE values for the response variable rating of $MR = 50\%$, indicating that the fiMI method has the best imputation effect. Overall, for goalkeeper data set with a large ratio of missing, the fiMI method has the highest imputation accuracy, followed by the EMRE and ERLS methods, respectively.

IV. Conclusion

Big data statistical analysis has become one of the mainstream positions in current statistical research. As missing data in statistical analysis is objective and inevitable in reality, techniques for dealing with missing data have received much attention from the statistical community, and imputation methods for missing data have been widely used in many fields. To address this issue, this paper investigates imputation methods for handling missing response variables in linear regression models to better improve the imputation accuracy while interpolating missing data. The work accomplished is as follows: the six methods are compared in terms of method steps, the advantages and disadvantages of the six methods are summarised, and the six methods are compared in terms of computational performance.

For the problem of imputation accuracy, the sensitivity and stability of the method are investigated through simulation, and real data analysis is carried out to verify the performance of the method. It is found that the proposed method has higher imputation accuracy and is more effective in dealing with data with missing ratios.

The performance of the imputation method discussed in this paper is mainly verified through a large number of simulation experiments and real data, mainly from practice and applications. Next, more attention will be paid to theoretical support, and the theorem proving of each imputation method will be studied, which can also serve as a direction for us in the future.

References

[1] J. X. Wang and M. Yu, “Note on the EM algorithm in linear regression model,” *International Mathematical Forum*, vol. 4, no. 38, pp. 1883-1889, 2009.

[2] S. V. Buuren and K. G. Oudshoorn, “mice: Multivariate imputation by chained equations in r,” *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011.

[3] G. Guo, Y. Sun and X. Jiang, “A partitioned quasi-likelihood for distributed statistical inference,”*Computational Statistics*, vol. 35, no. 4, pp. 1577–1596, 2020.

[4] G. Guo, H. Song, and L. Zhu, “ISR: The Iterated Score Regression-Based Estimation Algorithm,” 2022.

[5] G. Guo, C. Wei, and G. Qian, “Sparse online principal component analysis for parameter estimation in factor model,” *Computational Statistics*, vol. 38, no. 2, pp. 1095-1116. 2022.

[6] C. Chang, Y. Deng, X. Jiang and Q. Long, “Multiple imputation for analysis of incomplete data in distributed health data networks,” *Nature Communications*, vol. 11, no. 1, pp. 5467-547, 2020.

[7] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. John Wiley, 1999.

[8] J. L. Schafer, *Analysis of incomplete multivariate data*. London: Chapman &Hall, 1997.

[9] R. J. A. Little, “Missing-Data Adjustments in Large Surveys,” *Journal of Business & Economic Statistics*, vol. 6, no. 3, pp. 287-296, 1988.

[10] T. P. Morris, I. R. White and P. Royston, “Tuning multiple imputation by predictive mean matching and local residual draws,” *BMC Med Res Methodol*, vol. 14, no. 1, pp. 1-13, 2015.

[11] S. V. Buuren, *Flexible Imputation of Missing Data*, 2nd ed. Chapman & Hall/CRC, 2018.