International Journal of Mathematics and
Computer Science, **19**(2024), no. 4, 1189–1201

Missing Data Estimation Based on Iterated Penalized Linear Regressions Algorithms

Lorenzo J. Martinez H., David Gutierrez Duque

Natural and Exact Science Faculty
Caldas University
Manizales, Colombia

email: lorenzo.martinez_h@ucaldas.edu.co, davidgd2015@gmail.com

(Received April 7, 2024, Accepted May 8, 2024,
Published June 1, 2024)

Abstract

The presence of missing information is relatively frequent when several variables are observed in a group of individuals. In a given context, the missing information must be estimated; otherwise, the analysis, conclusions and corresponding recommendations of the information will be biased and unreliable. Therefore, a methodology for consistent and efficient estimating missing information must be properly chosen. Consequently, our aim is to develop an algorithm to estimate missing values by considering statistical techniques based on iterated penalized linear regressions.

1 Introduction

The presence of missing data when k variables are observed in a group of n individuals is a frequent problem in qualitative and quantitative studies and consecuently in analyses of univariate and multivariate data because this leads to biased and inconsistent estimations in the analysis. To avoid inconsistencies in the results obtained, different kind of missing data estimation techniques are used to obtain better recognition of the characteristics or attributes of interest. Currently, there are multiple algorithms and techniques

Key words and phrases: Missing data, estimation, linear regression.
AMS (MOS) Subject Classifications: 62D10, 62J05.
ISSN 1814-0432, 2024, http://ijmcs.future-in-tech.net

for dealing with missing data, one of the most common being listwise deletion which deletes all cases with missing values on one or more variables [5]. This technique is the least recommended since it reduces the sample size significantly [6]. Other techniques commonly used when dealing with missing data are those based on instransitive imputation algorithms. These algorithms do not take into account the other variables present in the dataset; for example, if the missing value lies in the variable X_j, the algorithm only considers data that is in the variable X_j, and does not consider data that is in the variable X_p, where $p \neq j$. Some popular algorithms are mean imputation and median imputation. To summarize, if x_{ij} is a missing value, its imputed value would be the mean or median of existent data of the variable X_j respectively. Intransitive imputation algorithms are especially valid when the correlation between the variables and the missing data percentaje are relatively low [8]. The most rigorous algorithm for estimation of missing data is the one where the estimation of a value that lies in the variable X_i depends on other variables X_p, where $p \neq j$. This is called transtitive imputation [6]. One of the most famous transitive imputation algorithms are those based on regression imputation. It works by replacing each missing value by an estimated value based on a regression model with the existent variables [7, 9]. This technique is relatively effective when the correlation among observed variables are stable which is relatively feasible to accomplish when the sample size is big [8]. These particularities are quite common in experimental results in the natural and social sciences. Its advantages are to avoid altering significantly the standard deviations and the distribution [7]. However, it generates inflation of the correlations and variances [9] and lack of variability in the imputed data [5]. These disadvantages are mainly due to dispensing with the imputation error [5, 9].

2 In the quest for a new algorithm

Considering the need for a more consistent applicable and rigorous algorithm, the following terms should be considered:

i) Estimated values for missing data must be based on all variables in the data set.

ii) The algorithm must be iterative in order to obtain more accurate estimated values.

iii) Each time the algorithm is iterated, the estimated value must converge and not diverge.

iv) Most sets of social and life science data sets follow a non-parametric distribution and so the algorithm must work properly with non-parametric data sets.

v) The estimated values must have an associated estimation error based on a given confidence level.

2.1 Linear regression

Below, the basic theoretical aspects of the linear regression model are briefly described. The mathematical model associated with these tests is theoretically supported by the following assumptions.

- Normality of residuals,
- Homoscedasticity or equality of variances among treatments, and
- Independence among observations.

2.1.1 Multiple linear regression

Regression analysis is the most commonly used statistical technique for investigating and **modeling the relationship between variables**. Its appeal and utility often stem from the logically sound process of using an equation to express the connection between a variable of interest (the response) and a set of related predictor variables.

2.1.2 Multiple linear regression model

Let $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_k$ be k predictor variables believed to be related to a response variable $\mathbf{Y}$. The classical multiple linear regression model suggests that the variable Y consists of a mean which continuously depends on the $\mathbf{X}_i'$s, and a random error ε, representing measurement errors and the effects of other variables not explicitly included in the model. The predictor variable values are treated as fixed, while the error and the response are regarded as random variables whose behavior is characterized by a set of distributional assumptions.

Thus the classical multiple linear regression model in matrix form can be written as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{2.2}$$

In general, $\mathbf{Y}$ is an $n \times 1$ matrix of observations, $\mathbf{X}$ is an $n \times (k+1)$ matrix of predictor variables, $\boldsymbol{\beta}$ is a $(k+1) \times 1$ vector of unknown regression coefficients, and $\boldsymbol{v}$ is an $n \times 1$ vector of random errors.

In order to establish the properties of the estimators obtained, we need to specify a set of hypotheses about the model (Montgomery et al., 2004).

The main assumptions about the regression model are as follows:

- **Linearity:** $\mathbf{Y}$ is related to $\mathbf{X}$ through the regression model given by equation (2.1).
- **Constant variance**: $\mathrm{Var}\,(\varepsilon_i) = \sigma^2$, for $i = 1, \dots, n$.
- **Independence:** $\mathrm{Cov}\,(\varepsilon_i, \varepsilon_j) = 0$, for $i \neq j$.
- **Normality:** $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are unobservable random variables that are normally distributed with $\mu = 0$ and constant σ^2. This can be expressed as:
- **Predictor variables:** The variables $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_k$ are algebraically linearly independent and are treated as fixed, meaning they are not random variables.

Estimating the Regression model involves assigning numerical values to the unknown parameters $\boldsymbol{\beta}$ based on the available sample information of the observable variables in the model. We will only consider two estimation methods:

- Ordinary Least Squares (OLS) method.

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{Y}\right)\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}, \quad \hat{\boldsymbol{\sigma}}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - (k+1)}. \tag{2.3}$$

- Maximum Likelihood (ML) method

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{Y}\right)\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}, \quad \hat{\boldsymbol{\sigma}}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n}. \tag{2.4}$$

2.2 Missing Data Handling Techniques

2.2.1 Case Deletion

This is the most common technique and involves omitting cases (individuals/objects) with missing data in their variables (Kang, 2013). If the Missing Completely at Random (MCAR) pattern holds, then this technique can produce unbiased estimates (Kang, 2013). One of its drawbacks is that it can significantly reduce the sample size (Mostafa, 2019).

2.2.2 Mean Imputation

This method replaces the missing data with the mean value of the variable which is calculated using the available data (Kang, 2013). It is particularly valid when missing data follows a MCAR pattern (Song y Shepperd, 2007), and when the correlation between variables is low and the percentage of missing data is also low (Donner, 1982).

2.2.3 Regression Imputation

This is an example of transitive imputation, where the imputed value of the target variable depends on other variables (Mostafa, 2019). It replaces each missing data point with an estimated value based on a regression model using the available variables (Song y Shepperd, 2007; Kang, 2013). The technique is relatively effective when the correlations among observed variables are stable which is achieved when the sample size is large (Donner, 1982). Unlike intransitive imputation methods like mean and median, it avoids significantly altering standard deviations or the distribution shape (Kang, 2013). However, it can lead to inflation of correlations and variances (Song y Shepperd, 2007).

2.2.4 Expectation-Maximization (EM)

EM is a general method for finding the maximum likelihood estimation of parameters under a certain distribution in an incomplete dataset (Song y Shepperd, 2007). It is based on predicting missing data through the iterative estimation

of parameters until reaching a convergence of maximum likelihood estimation (Song y Shepperd, 2007), meaning a stability of the system (Kang, 2013).

One of its assumptions is that incomplete cases follow a Missing at Random (MAR) mechanism rather than a MCAR mechanism (Song y Shepperd, 2007). Its drawbacks lie in the potentially long time it might take to converge which, besides being time-consuming, could underestimate the standard error of estimation (Kang, 2013).

2.2.5 Single Imputation

Single imputation is a technique that relies on imputing a single value for each missing data point (Song y Shepperd, 2007). This technique does not reflect the uncertainty in the estimation of the missing data (Song y Shepperd, 2007) which is addressed by multiple imputation.

2.2.6 Multiple Imputation (MI)

Multiple Imputation imputes missing values multiple times, aiming to capture the sampling variability of a complete dataset (Song y Shepperd, 2007; Kang, 2013). This technique incorporates the uncertainty arising from missing data (Kang, 2013). It has been found to be robust against violations of normality assumptions, small sample sizes, or a high percentage of missing data (Kang, 2013). Its limitations lie in the fact that each application of MI can produce different results, making the technique non-replicable (Song y Shepperd, 2007).

3 Methodology

In the presence of a dataset with missing information, the user must decide whether it is advisable to exclude individuals from the study who lack observations on one or more variables of interest, or instead retain them and proceed with the necessary estimations. The obtained results will play a pivotal role in decision-making and future recommendations.

The challenges associated with analyzing fragmented data are well recognized within the context of descriptive and inferential statistics. Examples include biases and inefficient estimations, among others. Therefore, there's a need to employ statistical methodologies that enable, in one way or another, consistent and efficient estimations of missing information. Consequently,

various estimation strategies are employed, including mean imputation, median imputation, regression-based imputation, etc.

To address these issues, the aim is to develop an algorithm, along with its corresponding documentation, based on iterated and penalized linear regressions. This algorithm will explore optimal conditions for its performance. To achieve the objectives set forth in this research, the following stages will be undertaken:

First Stage

Review the theoretical foundations of methodologies related to missing data estimation. It is important to clarify that this work is concerned with situations such as illustrated in the following table, where missing data occurs when k variables are observed in n individuals, NA indicating missing data (Not Available). The missing information to be considered for estimation will pertain to variables located at least on an ordinal measurement scale. Some methodologies for missing data estimation considered in this work and whose theoretical foundations will be subject to review are mentioned below, for example:

- Deletion methods,
- Imputation methods,
- Expectation-Maximization Algorithm.

Second Stage

Design and construction of an algorithm for missing data estimation in the R programming language, based on iterated linear regressions, along with its corresponding documentation.

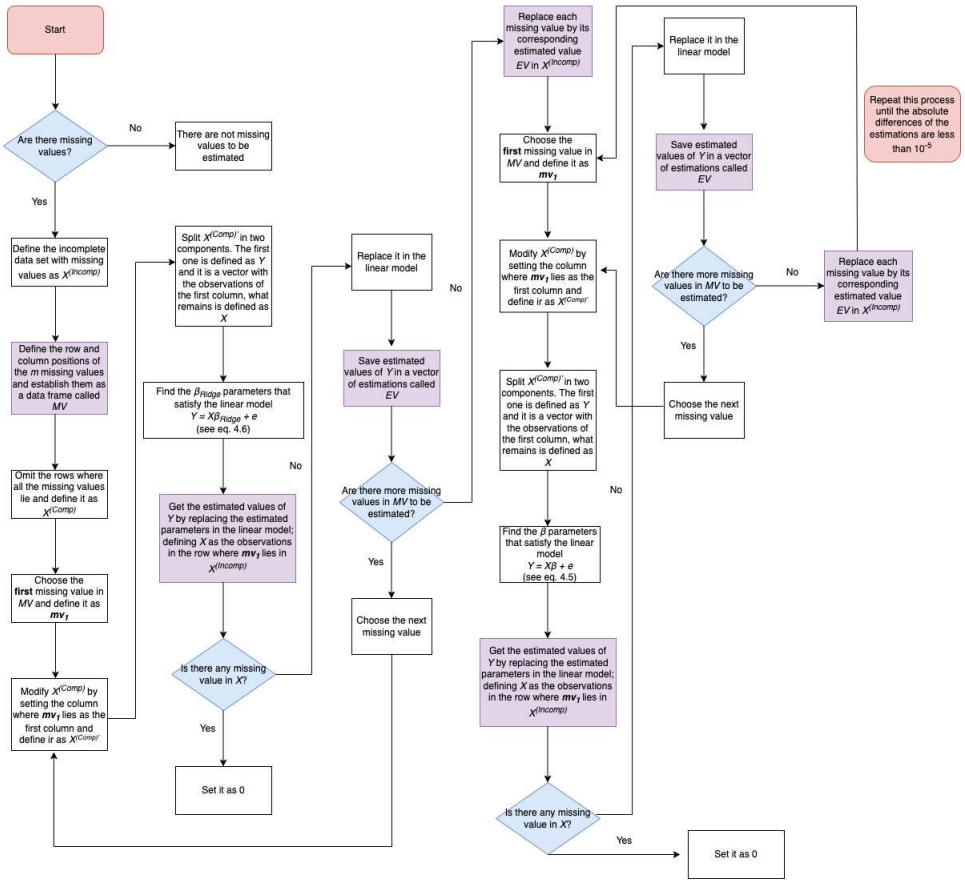

Figure 1: Flowchart of the proposed algorithm.

Third Stage

In order to compare the performance of existing estimation algorithms in the R software with the performance of the algorithm proposed in this work, simulations will be conducted using databases from different probability distributions. The behavior resulting from simulations will be explored, where one or several parameters associated with considered distributions are modified, allowing the identification of optimal performance conditions. The algorithm's behavior under assumption violations will also be examined, thus investigating its robustness. More precisely, individual or simultaneous parameter variations will help identify conditions for good performance. Performance tests will then be conducted in practical scenarios, involving datasets from R and databases from various application areas, where information is

missing and corresponding estimations are made. The proposed algorithm will also be subjected to situations involving multivariate data analysis or experimental design.

4 Results and Analysis

4.1 Statistical Package Used

The statistical algorithm was implemented as an R package (available at https://davidbiol.github.io/empire/index.html) which includes the following functions:

```
count_miss(data)
```

This function determines the number of missing data points, where data is a matrix or data frame of data.

```
pos_miss(data)
```

This function determines the row-column positions of missing data points, where "data" is a matrix or data frame with data.

```
impute_mean(data)
```

Intransitive imputation based on the mean of the variable where the missing data point is located, where "data" is a matrix or data frame with data.

```
impute_median(data)
```

Intransitive imputation based on the median of the variable where the missing data point is located, where "data" is a matrix or data frame with data.

```
estimate_mlr(data, diff)
```

Estimates missing data points using linear regressions that satisfy certain conditions, where "data" is a matrix or data frame with data.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}, \tag{4.5}$$

where $\mathbf{Y}$ is a vector of the variable with the missing data point and $\mathbf{X}$ corresponds to a matrix with the rest of the complete data.

The parameters data or data frame and diff correspond to the data matrix and the minimum difference desired between the estimated values in each iteration and their previous iteration, respectively.

```
estimate_ridge(data, diff, ridge_alpha)
```

Estimates missing data points using ridge regressions where data is the data matrix, diff is the desired minimum difference between estimated values in each iteration and their previous iteration, and ridge_alpha is the regularization parameter for ridge regression.

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{Y} \tag{4.6}$$

Here, $\mathbf{Y}$ is a vector of the variable with the missing data point, $\mathbf{X}$ corresponds to a matrix with the rest of the complete data and λ is the ridge parameter that corresponds to the penalty, which should be an integer greater than or equal to 0. A higher λ value results in a stronger penalty.

The parameters data or data frame correspond to the data; diff is the desired minimum difference between the estimated values in each iteration and their previous iteration; ridge_alpha corresponds to the penalizing parameter λ.

4.2 Monte Carlo Simulations

To determine the efficiency of the algorithm, Monte Carlo simulations were conducted, where the following assumptions were violated and/or met: Multivariate Normality, Multicollinearity, Homoscedasticity.

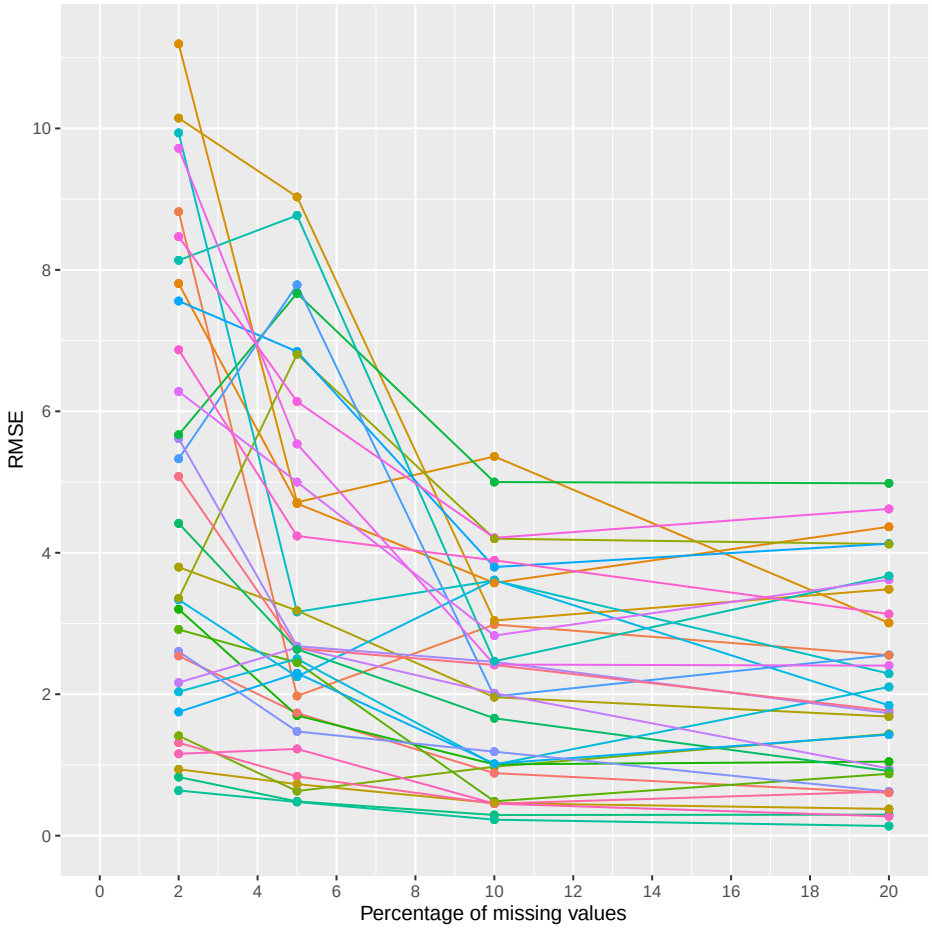

Figure 2: Root Mean Squared Error (RMSE) as a function of the percentage of missing data in the absence of multicollinearity.

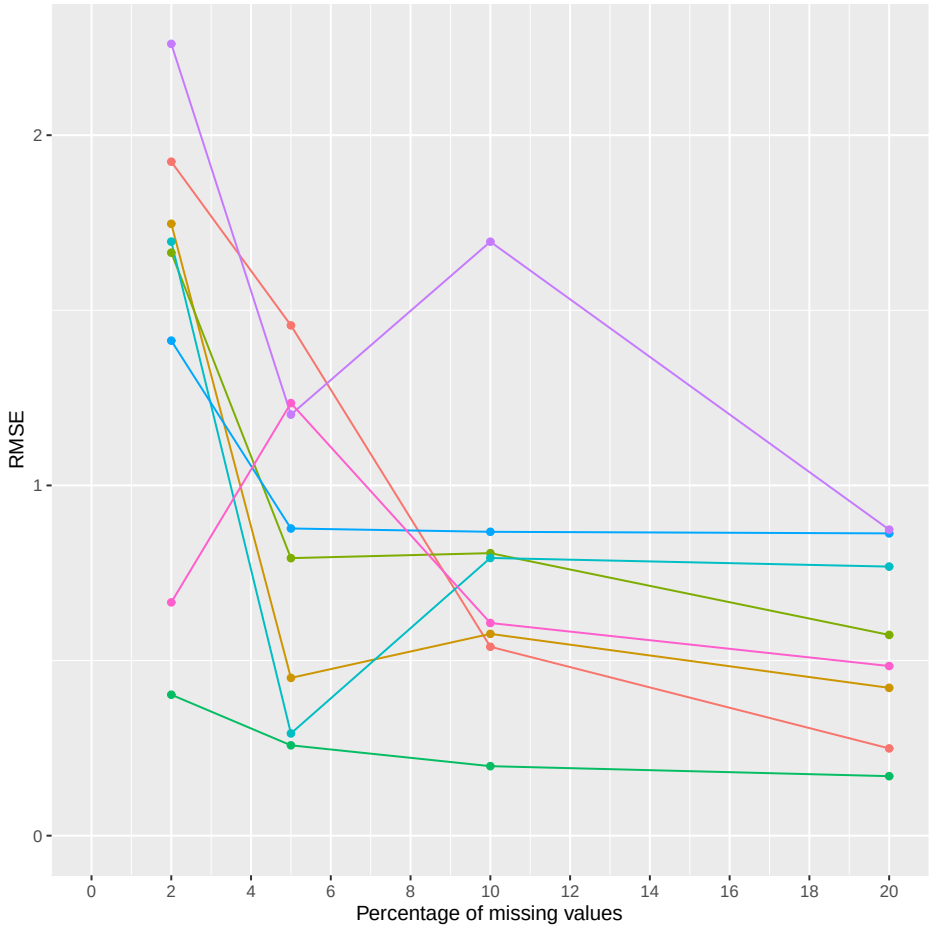

Figure 3: Root Mean Squared Error (RMSE) as a function of the percentage of missing data in the absence of multicollinearity and multivariate normality.

5 Conclusions

Empire is an algorithm that efficiently estimates missing data. However, it is important to understand the nature of the data to avoid incorrect estimation. This nature includes the distribution of variables, the percentage of missing data, the pattern of data loss, the mechanism of data loss, and the specific experimental context. Ridge penalization helps alleviate some constraints observed in linear regression.

The statistical package can be installed using the following commands:

```
install.packages("remotes")
remotes::install_github("davidbiol/empire")
```

References

[1] Handbook of Missing Data Methodology, Edited by Geert Molenberghs et al., Taylor & Francis Group, 2015.

[2] R. Little, D. Rubin, *Statistical Analysis with Missing Data,* Second Edition, Wilie Series in probability and statistics, 2002.

[3] P. McKnight, K. McKnight, S. Sidani, A. Figueredo, *Mising Data-A Gentle Introduction*, 2007.

[4] Estimación de datos faltantes con el Algoritmo EM. Tesis para obtener el título de Actuario, M. F. Lerdo, Universidad Autónoma de México, 2014.

[5] J. Peugh, C. Enders, Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement, Review of Educational Research, (2004), 74. 525-556.10.3102/00346543074004525.

[6] S.M. Mostafa, Imputing missing values using cumulative linear regression, CAAI Transactions on Intelligence Technology, (2019), 1–19.

[7] H. Kang, The prevention and handling of the missing data, Korean Journal of Anesthesiology, **64,** no. 5, (2013), 402–406.

[8] A. Donner, The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values, The American Statistician, **36,** no. 4, (1982), 378–381.

[9] Q. Song, M. Shepperd, Missing data imputation techniques, International Journal of Business Intelligence and Data Mining, **2,** no. 3, (2007), 261–291.