A New Algorithm to Impute the Missing Values in the Multivariate Case

Isaac Almasi [1]**, Mohsen Salehi** [2]**, and Mohammad Moradi** [1]**.**

[1] Department of Statistics, Faculty of Science, Razi University, Kermanshah, Iran.
[2] Department of Statistics, Faculty of Science, University of Qom, Qom, Iran.

Abstract. There are several methods to make inferences about the parameters of the sampling distribution when we encounter the missing values and the censored data. In this paper, through the order statistics and the projection theorem, a novel algorithm is proposed to impute the missing values in the multivariate case. Then, the performance of this method is investigated through the simulation studies. In an attempt to validate the proposed method and compare it with some other methods a real data is used.

Keywords. Order Statistic, Imputation, Projection Theorem, Censored Data, Linear Regression Model.

MSC: 62D05, 62N01.

1 Introduction

In many types of researches, some data values are missing or not recorded. In this situation, many software and the conventional statistical methods, usually are remove the missing values and use the usual analysis methods for incomplete data. Due to removing the missing values in the data set, the sample size will be reduced and make it difficult to have accurate estimation of the parameters of population distribution. In addition, it is possible that the characteristics of the missing values are different from other values, therefore the large bias occurs in the estimation of parameters Allison (2001). In general, three different assumptions make mechanism of the missing values:

Corresponding Author: Isaac Almasi (i.almasi@razi.ac.ir)
Mohsen Salehi (m.salehi@qom.ac.ir)
Mohammad Moradi (moradi_m@razi.ac.ir)

missing completely at random (MCAR), missing at random (MAR) and non-ignorable Rubin (1976). In the first one, using the incomplete data can produce the unbiased estimators for the parameters Little and Rubin (2019). In the second one, the estimator of parameters might be consistent with some conditions on types of variables, including the existence of only two categorical random variables in the data set Fleiss et al. (2013). But, in the third one, the third, none of the parameter estimations are consistent and any statistical inference requires some assumptions on the relationship between the variables with missing and non-missing values Kim and Yu (2011); Yuan and Yin (2010). The EM algorithm is also a public method to estimate the parameters in an incomplete data. Dempster et al. (1977) and McLachlan and Krishnan (2007) proposed a new algorithm based on EM algorithm that estimates the parameters of the nonlinear structural equations from the random missing data.

There are some other common methods, including imputation of the missing values by mean, median or mean of the nearest observed value, which try to keep the sample size and improve the accuracy of the estimated parameters. Although the mean imputation is the simplest imputation method, it is not attractive, because imputation of all missing values with the mean leads to the two problems; the random perturbations in the distribution of data and underestimation of variance Haitovsky (1968). The other imputation method is regression imputation. In this method, the missing values are replaced by the predicted values obtained by the regression of the variable containing missing values as a response variable, and the other variables as the independent variables Greenland and Finkle (1995); Gold and Bentler (2000); Musil et al. (2002).

In the recent decade, many types of researches have been done over the reconstruction of the order statistics from the observed data. In this regards, an upper bound for the mean of the order statistics is obtained Klimczak and Rychlik (2005). Then, through the inverse hazard rate and based on the left censored sample, the reconstruction of the past failure times is investigated for the proportional models Asgharzadeh et al. (2012). Then, some point and interval reconstruction are computed for the missing order statistics in the exponential distribution using the classic approach Razmkhah et al. (2010). This study was examined using the Bayes approach Khatib et al. (2013). A novel method based on the projection theorem, is proposed by Almasi et al. (2017) to reconstruct the missing order statistics.

Considering the missing at random in some values in the response of the linear regression model, we propose a new algorithm to reconstruct them based on the order statistics. The rest of the article is organized as follows. In Section 2, we describe the imputation method through the order statistics and propose the best reconstruction for the i-th order statistic in the random sample from the arbitrary distribution, especially the standard uniform distribution. In Section 3, the proposed imputation method is presented in the linear regression model. Then, two cases are considered for the underlying distribution of residuals in the linear regression model which contains the missing values in the response variable in Section 4. Moreover, the mean square error

of the proposed method and some other considered methods are compared through the simulation study. To validate the proposed method, we apply it to the real data in Section 5. Finally, the results are discussed in Section 6.

2 Imputation in the Order Statistics

Suppose that $Y_1, \ldots, Y_n$ denote a random sample from a population with cumulative distribution function (CDF), $F_Y(y)$, and $Y_{(1)}, \ldots, Y_{(n)}$ are its corresponding order statistics. Assume that some of the order statistics are censored or missed. Thus, the vector of observed order statistics is reduced to $(Y_{(1)}, \ldots, Y_{(r)}, Y_{(s)}, \ldots, Y_{(n)})$ for all $0 \le r < s < n + 1$. It is proved that the projection mapping of $Y_{(l)}$, $r < l < s$, onto $\mathcal{M}(Y_{(1)}, \ldots, Y_{(r)}, Y_{(s)}, \ldots, Y_{(n)})$ and $\mathcal{M}(Y_{(r)}, Y_{(s)})$ are the same Almasi et al. (2017). In addition, the best mean square predictor of $Y_{(l)}$ in $\mathcal{M}(Y_{(r)}, Y_{(s)})$, where $\mathcal{M}(Y_{(r)}, Y_{(s)}) = \{Z \in L^2 : Z = \phi(Y_{(r)}, Y_{(s)})$ for some $\phi : \mathbb{R}^n \to \mathbb{R}\}$, is the conditional expectation $E(Y_{(l)}|Y_{(r)}, Y_{(s)})$ Brockwell and Davis (1991). Therefore, considering $\mu_i = E(Y_{(i)}), \sigma^2 = Var(Y_{(i)})$, and $\rho_{i,j} = Corr(Y_{(i)}, Y_{(j)})$ for all $i, j = r, l, s$, the best linear reconstruction, prediction, of the l-th order statistic proposed by Almasi et al. (2017) is as follows:

$$\hat{Y}_{(l)} = a_0 + a_r Y_{(r)} + a_s Y_{(s)}; \quad 0 \le r < l < s < n + 1,$$

where,

$$a_r = \frac{\rho_{r,l} - \rho_{r,s}\rho_{l,s}}{1 - \rho_{r,s}^2} \cdot \frac{\sigma_l}{\sigma_r} \; ; \quad a_s = \frac{\rho_{l,s} - \rho_{r,s}\rho_{r,l}}{1 - \rho_{r,s}^2} \cdot \frac{\sigma_l}{\sigma_s} \; ; \quad a_0 = \mu_l - a_r\mu_r - a_s\mu_s. \tag{2.1}$$

The determination of a_0, a_r and a_s is usually time consuming and that's why we introduce an approximation of best linear reconstruction based on order statistics according to the standard uniform distribution. First, in the following, the best linear reconstruction of the l-th order statistic is obtained when the underlying distribution of variables is the standard uniform distribution.

Theorem 2.1. *Let $U_{(1)}, \ldots, U_{(r)}, U_{(s)}, \ldots, U_{(n)}$ be the order statistics according to the random sample from the standard uniform distribution, then the best reconstruction of l-th order statistic for all $0 \le r < l < s < n + 1$ is as follows:*

$$\hat{U}_{(l)} = U_{(r)} + \frac{l - r}{s - r}(U_{(s)} - U_{(r)}), \quad l = r+1, \ldots, s-1. \tag{2.2}$$

Proof. $U_{(i)}$ for all $1 \le i \le n$ is the i-th order statistic from the standard uniform distribution, then by Balakrishnan and Cohen (2014)

$$\mu_i = E(U_{(i)}) = \frac{i}{n + 1} \; ; \quad \sigma_i^2 = \frac{i(n + 1 - i)}{(n + 1)^2(n + 2)} \; ; \quad \rho_{i,j} = \left(\frac{i(n + 1 - j)}{j(n + 1 - i)}\right)^{\frac{1}{2}} \quad .$$

Therefore, the coefficients, a_0, a_r and a_s in equation (2.1) are computed as follows:

$$a_r = \frac{s - l}{s - r} \; ; \quad a_s = \frac{l - r}{s - r} \; ; \quad a_0 = 0.$$

Therefore, the theorem is proved. Considering $s = n + 1(U_{(n+1)} = 1)$ and $r = 0(U_{(0)} = 0)$, the right and left censors are accrued, respectively. □

Let $U_1, \ldots, U_n$ and $Y_1, \ldots, Y_n$ denote a random sample from a standard uniform distribution and distribution function, $F_Y(y)$, respectively. If the inverse distribution function F^{-1} is defined by $F^{-1}(u) = \sup\{x : F(x) \leq u\}$, then, for $i = 1, \ldots, n$, $Y_{(i)} \stackrel{d}{=} F^{-1}(U_{(i)})$, where $U_{(i)}$ and $Y_{(i)}$ are i-th order statistics in samples $U_1, \ldots, U_n$ and $Y_1, \ldots, Y_n$, respectively, Balakrishnan and Cohen (2014).

Remark 1. If $(y_{(1)}, \ldots, y_{(r)}, y_{(s)}, \ldots, y_{(n)})$ be the vector of order statistics from distribution function, $F_Y(y)$, from the above point, then $(F(y_{(1)}), \ldots, F(y_{(r)}), F(y_{(s)}), \ldots, F(y_{(n)}))$ is a realization of order statistics from standard uniform distribution. Thus, $F^{-1}(U_{(l)})$ may be considered as a reconstructor for $Y_{(l)}$ i.e., for $l = r + 1, \ldots, s - 1$,

$$Y_{(l)} = F^{-1}(U_{(l)}). \tag{2.3}$$

By substituting (2.2) into (2.3), we obtain approximated best linear reconstructor of the l-th order statistic in the distribution function, $F_Y(y)$, as follows:

$$\hat{Y}_{(l)} = F^{-1}(F(Y_{(r)}) + \frac{l - r}{s - r}(F(Y_{(s)}) - F(Y_{(r)}))) \quad l = r + 1, \ldots, s - 1, \tag{2.4}$$

That the given results corresponds to the unbiased conditional reconstructor introduced by Razmkhah et al. (2010).

3 The Introduced Algorithm for Imputation Missing Data

The regression method is a conventional method for imputing. Let

$$Y_j = \beta_0 + \sum_{i=1}^{k} \beta_j x_{ij} + \epsilon_j, \quad j = 1, 2, \ldots, n, \tag{3.1}$$

where Y_j and x_{ij} are the response variable and j-th observation of the i-th predictor, respectively. Suppose that there are no missing values in the independent variables, but there are m missing values in the response variable, Y. Assume that the missing values follows the missing at random mechanism Rubin (1976). Without loss of generality, let $\boldsymbol{y}_{obs} = (y_1, y_2, \ldots, y_m)$ and $\boldsymbol{y}_{mis} = (y_{m+1}, y_{m+2}, \ldots, y_n)$ are the observation and missing values for the response variable, respectively. In regression method, missing values are imputed as follows,

$$\hat{Y}_j = \beta_0 + \sum_{i=1}^{k} \beta_j x_{ij}, \quad for \quad j = m + 1, \ldots, n.$$

In our method, a missing value is imputed k times, based on each auxiliary variable, separately. Then a weighting average from the given values is computed.

Let $X = (x_{ij})$, $i = 1, 2, \ldots, k$, and $j = 1, 2, \ldots, n$ be the $n \times k$ matrix, without missing values of the observations of the independent variables. Denote $R = (r_{ij})$ as the $n \times k$ rank matrix where r_{ij} is the rank of the observation x_{ij} in the i-th column of X. The introduced algorithm includes the following steps.

Step 1. The coefficient correlation of y_{obs} and x_i, called r_i^*, is computed.

Step 2. Through (2.4), the estimation of y_j based on orders given from x_i orders given from x_i, which is denoted as y_{ij}^*, is computed as follows:

$$
y_{ij}^* = \left\{
\begin{array}{ll}
F^{-1}(\frac{l}{2}F(y_{(1)})) & \text{if } r_{ij} = 1, \\
F^{-1}(\frac{1}{2}(F(y_{(r_{ij}-1)}) + F(y_{(r_{ij})}))) & \text{if } r_{ij} = 2, 3, \ldots, m, \\
F^{-1}(\frac{r_{ij}-m}{n-m+1} + \frac{n-r_{ij}+1}{n-m+1}F(y_{(m)})) & \text{if } r_{ij} = m+1, \ldots, n,
\end{array}
\right\},
\tag{3.2}
$$

where $y_{(j)}$ is the j-th order statistic from y_{obs} and r_{ij} is determined as follows:

$$
r_{ij} = \left\{
\begin{array}{ll}
r_{ij} & \text{if } r_i^* > 0, \\
n + 1 - r_{ij} & \text{if } r_i^* < 0.
\end{array}
\right.
\tag{3.3}
$$

Step 3. The estimation of elements of y_{mis} are computed as follows:

$$
y_j^* = \sum_{i=1}^{k} w_i y_{ij}^*, \quad j = m+1, m+2, \ldots, n,
\tag{3.4}
$$

where $w_i = \frac{r_i^{*2}}{\sum_{i=1}^{k} r_i^{*2}}$.

4 Simulation Study

In this section, we consider two scenarios for the underlying distribution of the random error component in the linear regression model (3.1); normal and Cauchy distributions. Considering the missing values in the response variable, first, we compare the performance of the proposed imputation method (PI), with three other methods; the non-missing data (NM), the mean imputation (MI) and the conditional mean imputation (CMI), through a Monte Carlo simulation. Our objective is to compare mean square errors (MSEs). The MSE is averaging $\frac{1}{n-m}\sum_{j=m+1}^{n}(y_j - y_j^*)^2$ over all $N = 10000$ simulation iterations. Without loss of generality, we take $n = 12, 16$, and consider that the half of the y values are missing. In the first scenario, we consider two cases for the linear regression; the simple linear regression and the multiple linear regression model, $k = 2$. We set $E(\epsilon_i) = 0$ and $var(\epsilon_i) = \sigma^2$, for all $i = 1, 2, \ldots n$. In Case 1, ρ is taken from $\{0, \pm 0.01, \ldots, \pm 0.99\}$. The random sample, $n = 12$, are then generated from the bivariate normal with parameters $\mu_1 = 0, \mu_2 = 0, \sigma_1 = 1$ and $\sigma_2 = 1$.

The results shown in Figure 1 (*I*), demonstrates that the PI method has a less MSE compared to MI and CMI methods, when $0.4 \leq |\rho| \leq 0.8$. In this situation, when there

is low correlation between the response and the independent variable, $|\rho| < 0.4$, the PI and CMI method have poor performance than the MI. In the contrary, the regression method has more performance than the others while there is a high correlation between these two variables, $0.8 < |\rho| < 1$.

Figure 1(*II*) illustrates the same results when $n = 16$ and $m = 8$. In Figure 1(*III*), we investigate how much the estimation of parameters μ, σ, and ρ through the mentioned methods are close to its assumed values. For this purpose, after the imputation of the missing values through the proposed and other mentioned methods, the parameters μ, σ, and ρ are estimated again. Subfigure 1(*III*)A shows that the estimated means are around zero, which is close to $\mu = 0$ in all methods. Subfigure 1(*III*)B depicts that the estimate of parameter σ is closer to the initial value of $\sigma = 1$ through the PI method than the other methods when $0 < \rho < 0.65$. The same is true for the CMI method when $0.65 \leq \rho < 1$. In this Subfigure, we can see that the estimation of the mentioned parameter is not suitable for the MI method. Subfigure 1(*III*)C demonstrates that the CMI method presents the best estimate for ρ, whereas the PI has an acceptable performance for $0.7 \leq \rho < 1$. This method overestimate the parameter when $0 < \rho < 0.7$.
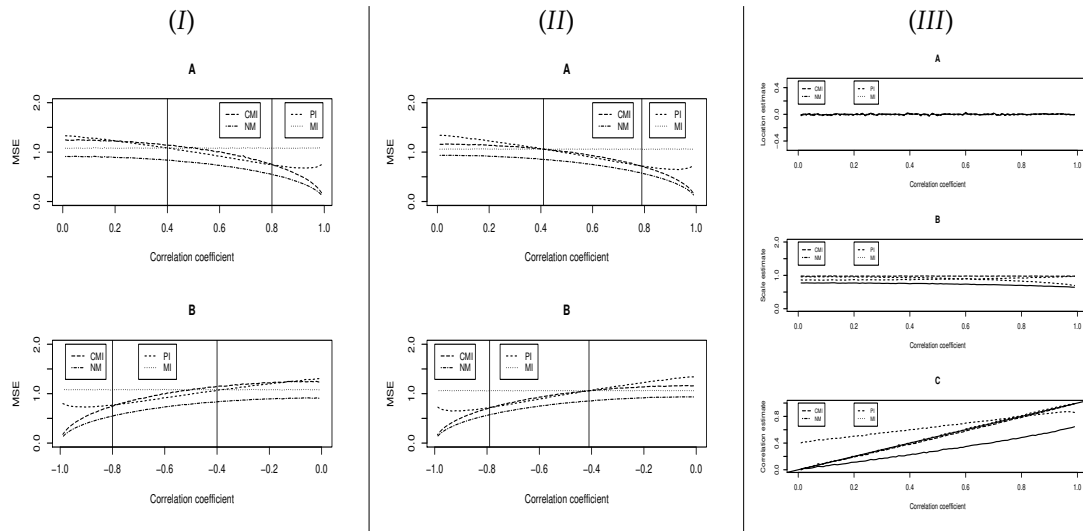

Figure 1: Imputation methods, PI, MI, CMI and NM in the simple linear regression model. The underlying distribution of the residuals is considered the normal distribution with the location and scale parameters, $\mu = 0$, and $\sigma = 1$, respectively. Subfigures (*I*) and (*II*) show the MSE comparison through these methods when $n = 12, m = 6$ and $n = 16, m = 8$, respectively and $|\rho| \leq 1$, and Subfigure (*III*) shows the validation of the mentioned methods when $n = 12, m = 6, |\rho| \leq 1$.

In Case 2, the random vector (Y, X_1, X_2) is generated from the standard trivariate normal distribution with parameters ρ_1, ρ_2 and ρ_3, which the first two parameters are the correlation between y and x_i for $i = 1, 2$, and the last one is the correlation between x_1 and x_2. For the minimum correlation between x_1 and x_2, we set $\rho_3 = 0.1$.

We consider various values for ρ_i, $i = 1, 2$, such that $\rho_2 = 0.1\rho_1$ to $\rho_2 = 0.3\rho_1$, and set ρ_1 from .01 to .99 by increments of 0.1. In addition, the simulation is replicated $N = 10000$ times. Then, the mean square errors (MSE) are computed for the proposed method and other mentioned methods in each iteration. Figure 2(*I*) depicts the results. We can see that the PI method has less MSE than the MI and CMI methods when $0.35 \leq \rho_1 < 0.85$. In addition, the regression and proposed methods do not have acceptable MSE, whereas the MI and regression methods have the minimum MSE while $\rho_1 < 0.35$ and $0.85 \leq \rho_1 \leq 1$, respectively. Then, we set $\rho_2 = 0.4\rho_1$ to $\rho_2 = 0.6\rho_1$. Take $\rho_1 \in \{0, \pm0.01, \ldots, \pm0.99\}$. Figure 2(*II*) shows that the PI method has the minimum MSE when $0.29 \leq \rho_1 \leq 0.81$, whereas the MI and regression methods have the minimum MSE for $\rho_1 < 0.29$ and $0.81 < \rho_1 \leq 1$, respectively. Then, we set $\rho_2 = 0.7\rho_1$ to $\rho_2 = 0.9\rho_1$, and ρ_1 from .01 to .99 by increments of 0.1. Figure 2(*III*) shows that the PI method has the minimum MSE when $0.35 \leq \rho_1 \leq 0.85$, whereas the MI and regression methods have the minimum MSE while $\rho_1 < 0.35$ and $0.85 < \rho_1 \leq 1$, respectively.
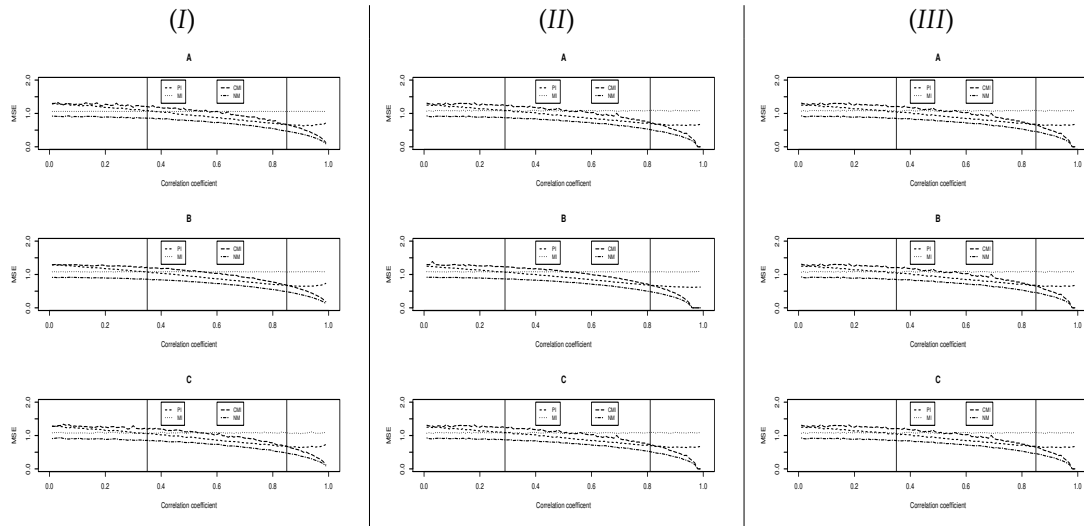

Figure 2: Imputation methods, PI, MI, CMI and NM in the multiple linear regression model, $k = 2$ when $n = 12, m = 6$, and $\rho_1 \in \{0.1, 0.11, \ldots, 0.99\}$. Subfigures (*I*)A, (*I*)B, and (*I*)C are correspond to $\rho_2 = 0.1\rho_1, \rho_2 = 0.2\rho_1$, and $\rho_2 = 0.3\rho_1$, respectively. Subfigures (*II*)A, (*II*)B, and (*II*)C are correspond to $\rho_2 = 0.4\rho_1, \rho_2 = 0.5\rho_1$, and $\rho_2 = 0.6\rho_1$, respectively. Subfigures (*III*)A, (*III*)B, and (*III*)C are correspond to $\rho_2 = 0.7\rho_1, \rho_8 = 0.2\rho_1$, and $\rho_2 = 0.9\rho_1$, respectively.

In the second scenario, it is assumed that the underlying distribution of ϵ_i is Cauchy distribution in the linear regression model (3.1). In order to reconstruct the missing values, we use the median imputation (MeI), and apply the PI and other mentioned methods in the linear regression model. In this regard, the vector (Y, X_1, X_2) is generated by the Heavy Package in R software. By Heavy Package, we can obtain prediction for missing values that indicated in results by the Heavy imputation (HI). We set $\rho_1 = \pm\rho_2$ taken from $\{0.1, 0.11, \ldots, 0.99\}$ and $\rho_3 = 0.1$. Then, the six values, $m = 6$, are removed

from the response variable and replicated the simulation $N = 10000$ times. Finally, MSEs are computed for all 10000 replications. Figure 3 (*I*) illustrates the results. We can see that the PI method, as well as the NM method, have less MSEs compared to the other methods when $0.35 \leq |\rho| \leq 0.75$. When $|\rho| < 0.35$ and $0.75 < |\rho| \leq 1$, the MI and the regression method (HI) has the minimum MSE, respectively.

In Figure 3(*II*), we check the accuracy of the estimation of parameters μ, σ, and ρ in the mentioned methods comparing to their assumed values. For this purpose, after the imputation of the missing values through the all mentioned methods, the parameters μ, σ, and ρ are estimated again. Subfigures (*II*)A shows that the estimation of mean is about zero and it is close to $\mu = 0$ in the PI and HI methods. But, the MeI presents a biased estimation for $\mu = 0$. Subfigure (*II*)B shows that the estimation of parameter σ is closer to the assumed value of $\sigma = 1$ through the PI and HI methods than the MeI methods. In this Subfigure, the MeI method does not present a suitable estimation for the mentioned parameter. In this regards, Subfigure C demonstrates that the HI method provides the best estimation for ρ, whereas the PI has an acceptable estimation for it when $0.6 \leq \rho < 1$. This method overestimates it when $0 < \rho < 0.6$.
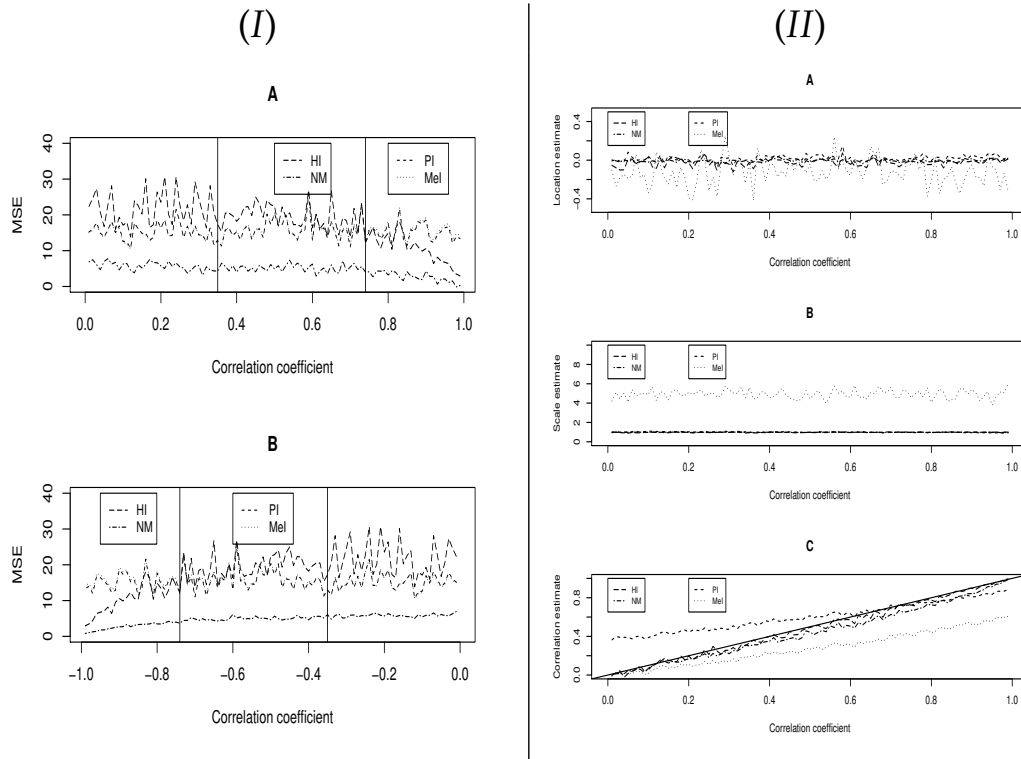

Figure 3: Imputation methods, PI, MeI, HI and NM in the simple linear regression model. The underlying distribution of the residuals is considered the standard Cauchy distribution. Subfigures (*I*) shows the MSE comparison through these methods when $n = 12, m = 6$ and, $|\rho| \leq 1$. Subfigure (*II*) depicts the validation of the mentioned methods.

5 Real Data

The real data is about the relationship between the percent of fruit worm, y, and the number of fruits, x, on the twelve trees (Table 6.9.1 Snedecor and Cochran (1967)). The correlation between these variables is -0.88. Considering the first four values in the response variable as the missing values, we apply PI and CMI methods on the data. Then, this carries out for the second and third four values of the response variable. Table 1 contains the results. The bold values show the better predictions. Although the correlation value is out of the suitable bound for the correlation value, $0.4 < |\rho| < 0.7$, we can see that the PI method is reconstructing the missing values much closer to the initial values of the response except in the second value of the response.

Table 1: Performance of the PI and MCI methods on the relationship between the fruit worm and the fruits number of the trees

Case number	y	x	MCI	PI
1	59.0	8.0	55.36	**55.27**
2	58.0	6.0	**58.23**	61.45
3	56.0	11.0	52.58	**52.47**
4	53.0	22.0	**40.82**	40.85
5	50.0	14.0	**50.07**	50.54
6	45.0	17.0	**47.22**	48.52
7	43.0	18.0	**47.22**	48.52
8	42.0	24.0	**39.67**	36.33
9	39.0	19.0	45.54	**44.71**
10	38.0	23.0	41.27	**39.71**
11	30.0	26.0	39.15	**34.99**
12	27.0	40.0	**19.55**	30.49
MSE			32.24	**25.92**

6 Conclusion

The missing values might be an obstacle in analyzing any data. In this article, we proposed a new imputation method that can be applied to analyze the linear regression model with the missing values in the response variable. This method is based on imputation using the order statistics. We present this method based on two assumptions on the distribution of error component, normal and Cauchy distributions, in the linear regression model. Through the simulation studies, we show that the proposed method has a better performance than the conventional methods mentioned in this article when the underlying distribution of the error component is normal and $0.4 < |\rho| < 0.7$. In addition, we show that the proposed method has a better performance while the underlying distribution is Cauchy and $0.3 < |\rho| < 0.8$.

Acknowledgments

References

Allison, P. D. (2001), *Missing data* (Vol. 136). Sage publications.

Almasi, I., Mohammadpour, A. and Mohammadi, M. (2017), Best linear unbiased interpolation of order statistics. *Communications in Statistics-Simulation and Computation*, **46**(5), 4161–4171.

Asgharzadeh, A., Ahmadi, J., Mirzazadeh Ganji, Z. and Valiollahi, R. (2012), Reconstruction of the past failure times for the proportional reversed hazard rate model. *Journal of Statistical Computation and Simulation*, **82**(3), 475–489.

Balakrishnan, N. and Cohen, A. C. (2014), *Order statistics and inference: estimation methods*. Elsevier.

Brockwell, P. J. and Davis, R. A. (1991), *Time series: theory and methods*. second ed., Springer Series in Statistics, Springer-Verlag, New York.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical society*. Series B (methodological), 1–38.

Fleiss, J. L., Levin, B. and Paik, M. C. (2013)], *Statistical methods for rates and proportions*. John Wiley and Sons.

Gold, M. S. and Bentler, P. M. (2000), Treatments of missing data: A Monte Carlo comparison of rbhdi, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling*, **7**(3), 319–355.

Greenland, S. and Finkle, W. D. (1995), A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, **142**(12), 1255–1264.

Haitovsky, Y. (1968), *Missing data in regression analysis. Journal of the Royal Statistical Society*. Series B (Methodological), 67–82.

Khatib, B., Ahmadi, J. and Razmkhah, M. (2013), Bayesian reconstruction of the missing failure times in exponential distribution. *Journal of Statistical Computation and Simulation*, **83**(3), 501–517.

Kim, J. K. and Yu, C. L. (2011), A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association*, **106**(493), 157–165.

Klimczak, M. and Rychlik, T. (2005), Reconstruction of previous failure times and records. *Metrika*, **61**(3), 277–290.

Little, R. J. and Rubin, D. B. (2019), *Statistical analysis with missing data (Vol. 793)*. John Wiley and Sons.

McLachlan, G. and Krishnan, T. (2007), *The EM algorithm and extensions (Vol. 382)*. John Wiley and Sons.

Musil, C. M., Warner, C. B., Yobas, P. K. and Jones, S. L. (2002), A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research*, **24**(7), 815–829.

Razmkhah, M., Khatib, B. and Ahmadi, J. (2010), Reconstruction of order statistics in exponential distribution. *Journal of the Iranian Statistical Society*, **9**, 21-40.

Rubin, D. B. (1976), Inference and missing data. *Biometrika*, **63**(3), 581–592.

Snedecor, G. and Cochran, W. G. (1967), *Statistical Methods*, 6th, ed. Ames: Iowa State University Press.

Yuan, Y. and Yin, G. (2010), Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics*, **66**(1), 105–114.