A New Robust Imputation Method for Longitudinal Data with Non-Normal Continuous Outcomes

Nesma M. Darwish[2], Yasmin A. Mohamed[1], Ahmed M. Gad[1,3,*], Abdelnaser S. Abdrabou[1] and Wafaa M. Ibrahim[1]

[1]*Statistics Department, Faculty of Economics and Political Science, Cairo University, Egypt*

[2]*Faculty of Political Science, Economics and Business Administration, May University, Egypt*

[3]*Business Administration Department, Faculty of Business Administration, Economics and Political Science, The British University in Egypt (BUE), Egypt*

Abstract: Missing values is very common in longitudinal data and it is the main challenge in analysis of longitudinal data. Missing values have a significant effect on longitudinal data analysis because they lead to loss of information, biased estimates, and misleading results. In practice there is a need for an imputation method to deal with missing values.

Aim: In this study a new robust regression-based imputation method to deal with missing values in longitudinal data is proposed. This method utilizes the modified adaptive linear regression model and does not require the normality of the responses. It is a novel robust imputation method as it is introduced for the first time.

Results and Conclusion: The simulation results show that the proposed method performs well compared to other methods especially for multivariate t-distribution and Chi-square distribution. Also, the proposed approach is effective apart from the missingness rate.

Keywords: Longitudinal data, Missing values, Robust imputation, Single imputation.

1. INTRODUCTION

Longitudinal data is very common in many fields, such as medicine, public health, education, economics, biology, and more. Missing data is a great challenge in analyzing longitudinal data. Missing data leads to efficiency loss and biased results. Missing values is said to be dropout when a subject leaves the study prematurely, or otherwise it is intermittent. There are three types of missingness mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing not at Random [10]. In this article, the focus is on dropout pattern and MAR mechanism.

Imputation methods are introduced to handle missing values. The imputation methods can be classified as either single-imputation methods (SI) or multiple-imputation methods (MI). The chosen imputation method depends on two main assumptions. The first assumption is the missing data mechanism. Most of the available techniques assume MAR and MCAR [1]. The second assumption is the distributional assumptions of the responses. Most of the available methods assume that the responses are normally distributed. However, these assumptions are often violated in practice, where repeated measurements on the same subject can exhibit heavy tails, skewness, or extreme values. Few imputation methods are available in literature if the normality assumption is violated.

*Address correspondence to this author at the Statistics Department, Faculty of Economics and Political Science, Cairo University, Egypt; Email: ahmed.gad@feps.edu.eg

The aim of this paper is to introduce a new robust imputation method to handle longitudinal data with missing values and compare it with the available single imputation methods using a simulation study. The proposed method is the Modified Adaptive Linear Regression (MALR) estimator that suggested by [3]. It offers an effective alternative by merging the robustness of the Modified Least Absolute Deviation (MLAD) estimator with the effectiveness of the Generalized Least Squares (GLS) approach. With this combination, the MALR-based imputation method can adjust to the data's underlying distribution and can accommodate heavy-tailed or non-normal distribution.

The rest of the article is organized as follows. Section 2 introduces the used notation throughout the article. Section 3 presents the most common single imputation methods for handling missing data. Section 4 introduce the proposed robust imputation approach. In section 5, simulation studies are conducted to evaluate the performance of the proposed technique. The proposed approach is applied to real data in Section 6. Finally, concluding remarks and future research points are provided in Section 7.

2. NOTATIONS

Let Y_{ij} represents the response of subject i at time point j; for $i = 1, \dots, n$ and $j = 1, \dots, m_i$, where and X_{ij}a vector of p covariates observed at time t_{ij}, for $i = 1, \dots, n$ subjects, and for each subject, the responses and the covariates are intended to be measured at times $j = 1, \dots, m_i$. The total number of observations $N = \sum m_i$. They_{ij}denotes the value of the variable Y_{ij} and x_{ijk} denotes the value of X_{ijk}

recorded at time t_{ij} ($i = 1, \dots, n$, $j = 1, \dots, m_i, k = 1, \dots, p$). the vector $Y_i = [y_{i1} y_{i2} \dots \dots y_{im_i}]^T$ is referred to as the complete data. Sometimes this vector Y_i contain missing values. So, Y_i can be partitioned into two components; Y_i^o denotes the vector of observed responses on the ith subject and Y_i^m denotes the vector of missing responses on the ith subject. Let R_i be an m_i × 1 vector of response indicators, $R_i = (r_{i1} r_{i2} \dots \dots r_{im_i})^T$,with $r_{ij} = 1$ if y_{ij} is missing and $r_{ij} = 0$ if y_{ij} is observed.

3. SINGLE IMPUTATION METHODS

Imputation methods are the most used method for handling missing data. It is the process of replacing missing values with estimates based on the available data, then analyzing the "pseudo" complete data using standard methods. In single imputation methods each missing value is replaced by a single value. There are different methods of single imputation available in literature. These include, but are not limited to, the mean, the median, the last observation carried forward, the Hot Deck, and regression methods.

3.1. The Mean Imputation Method

In this method, the missing values are replaced by the mean value of the observed values. This method helps maintain the overall dataset size and ensures that missing data does not disrupt the analysis. However, ithas significant drawbacks; it reduces variability and can lead to biased estimates. Despite its limitations, the mean imputation is often used in exploratory data analysis or when missing data is minimal, as it provides a quick and easy solution for handling incomplete datasets [6].

3.2. The Median Imputation Method

In the median imputation the missing values are replaced with the median of the observed data. For datasets with outliers or skewed distributions, this approach can be particularly helpful. However, a limitation of median imputation is that it does not preserve relationships between variables, potentially leading to biased statistical analyses [11].

3.3. The Last Observation Carried Forward (LOCF)

The missing values are replaced by the last observed value of the subject of interest. This method makes an unrealistic assumption that the individual's observation has not changed at all since the last measured observation [7].

3.4. Hot Deck (HOT) Method

In this method each missing value is replaced with an observed value from a similar subject. The subject with missing value is called the "recipient", and the subject from which the value is used for imputation is called the "donor". In the hot-deck, the responding and non-responding subjects are classified based on covariates, and then imputation is performed by randomly selecting a donor to replace each non-respondent [5].

3.5. The Regression Imputation (RM) Method

In this approach, a regression model is built using observed data, where the variable with missing values is treated as the dependent variable, and the other available variables serve as independent predictors. Then, the model is used to predict the missing values based on the observed values. Regression imputation can provide more accurate estimates compared to simpler methods like mean or median imputation [8].

All the above methods are not robust methods. Up to our knowledge, no single robust imputation method has been suggested for longitudinal data.

4. THE PROPOSED IMPUTATION METHOD USING THE MALR MODEL

The proposed robust single imputation method depends on the modified adaptive linear regression method(MALR) that has been introduced in [3]. The MALR is able to handle heavy-tailed longitudinal data. It is a linear combination of the generalized least squares method (GLS) and the modified least absolute deviation method (MLAD). They argue that the MALR estimator is efficient for heavy-tailed distributions.

4.1. The Modified Adaptive Linear Regression Estimator

The LAD estimator is the estimator that minimizes least absolute errors. For the marginal model,

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \text{for } i = 1, \dots, n, \quad (1)$$

where $\boldsymbol{y}_i = [y_{i1} y_{i2} \dots \dots y_{\text{im}}]^T$ is an $m_i \times 1$ vector of observations for i^{th} subject, $\boldsymbol{X}_i$ is a $m_i \times P$ matrix of covariates, and $\boldsymbol{\beta}$ is a $p \times 1$ column vector of the parameters. It is assumed that $\boldsymbol{\varepsilon}_i \sim f(0, \boldsymbol{V}_i)$, and f is a specified multivariate distribution may be normal distribution or not; $\boldsymbol{\varepsilon}_i's$ are identically in distribution. $\boldsymbol{V}_i$ is m_i × m_i variance covariance matrix of $\boldsymbol{\varepsilon}_i$. The LAD estimator for $\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{LAD}}$,can be obtained by minimizing the function $\sum_{i=1}^{n} |\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}|$.

[3] suggest modifying the LAD estimator, to accommodate longitudinal data, depending on de-correlating data. This is done by multiplying the whole data by $\boldsymbol{V}_i^{-\frac{1}{2}}$. The decorrelated model is

$$\boldsymbol{V}_i^{-\frac{1}{2}}\boldsymbol{y}_i = \boldsymbol{V}_i^{-\frac{1}{2}}\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{V}_i^{-\frac{1}{2}}\boldsymbol{\varepsilon}_i .$$

Or equivalently it can be written as

$$\boldsymbol{y}_i^* = \boldsymbol{X}_i^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i^*,$$

where $\boldsymbol{y}_i^* = \boldsymbol{V}_i^{-\frac{1}{2}}\boldsymbol{y}_i$, $\boldsymbol{X}_i^* = \boldsymbol{V}_i^{-\frac{1}{2}}\boldsymbol{X}_i$, and $\boldsymbol{\varepsilon}_i^* = \boldsymbol{V}_i^{-\frac{1}{2}}\boldsymbol{\varepsilon}_i$. By applying the least absolute deviation (LAD) estimator on transformed data, the modified LAD (MLAD) for$\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MLAD}}$, can be obtained by minimizing the function $\sum_{i=1}^{n}|\boldsymbol{y}_i^* - \boldsymbol{X}_i^*\boldsymbol{\beta}|$.

In terms of the original $\boldsymbol{y}$and $\boldsymbol{X}$the MLAD can be obtained by minimizing the function $\sum_{i=1}^{n}\left|\boldsymbol{V}_i^{-\frac{1}{2}}\boldsymbol{y}_i - \boldsymbol{V}_i^{-\frac{1}{2}}\boldsymbol{X}_i\boldsymbol{\beta}\right|$.

[3] propose the modified adaptive linear regression (MALR) estimator to be a linear combination of the modified least absolute deviation (MLAD) estimator and the generalized least square estimator (GLS). The correlation structure of the longitudinal data is taken into account by both the MLAD and the GLS estimators. The form of the MALR estimator is

$$\widehat{\boldsymbol{\beta}}_{MALR} = \omega\widehat{\boldsymbol{\beta}}_{GLS} + (1-\omega)\widehat{\boldsymbol{\beta}}_{MLAD}, \quad (2)$$

where ω is the weight that reflects the nature of the error distribution. The value of ω can be obtained as described in [3].

The proposed MALR-based imputation method achieves robustness and efficiency by adaptively combining two estimators with complementary strengths. The Generalized Least Squares (GLS) estimator provides efficient estimates when the underlying data are approximately normal, as it minimizes the variance of the estimates. However, the GLS estimator is sensitive to outliers and heavy-tailed distributions, which can severely affect the results. In contrast, the Modified Least Absolute Deviation (MLAD) estimator minimizes absolute errors rather than squared errors, making it less sensitive to outliers or non-normality.

The MALR estimator is a weighted linear combination of the GLS and the MLAD estimators. The adaptive weight ω determines the relative contribution of each estimator depending on the shape of the error distribution. When the data exhibit normal behavior, ωassigns more importance to the GLS component, preserving efficiency. When the data are non-normal or contaminated with outliers, ωincreases the influence of the MLAD component, enhancing robustness.Through this adaptive mechanism, the proposed MALR imputation-based method maintains the desirable statistical efficiency of GLS under normal distribution while inheriting the robustness properties of MLAD under non-normal or heavy-tailed distributions. This feature allows the method to produce reliable imputations across a broad range of practical data scenarios commonly encountered in practical problems.

4.2. The Proposed Robust Imputation Method

The proposed single imputation method depends on the modified adaptive linear regression (MALR) estimator. The proposed MALR imputation method does not assume normality and is not sensitive to extreme values. The steps of the proposed MALR imputation method are listed below.

Step 1: Estimate the generalized least squares (GLS) parameters using all available (non-missing) data. The GLS accounts for the within-subject correlation in longitudinal data and serves as an efficient estimator when the data are approximately normal.

Step 2: Estimate the modified least absolute deviation (MLAD) parameters for model in Eq. (1). These estimates are robust to outliers and heavy-tailed distributions.

Step 3: Obtain the MALR estimator in Eq. (2).

Step 4: The missing values are predicted, using the obtained MALR estimates, depending on the relationship between the variable with missing values and the fully observed covariates. For any subject i with missing responses, the missing component $\boldsymbol{y}_i^m$ is imputed as:

$$\hat{\boldsymbol{y}}_i^m = \boldsymbol{X}_i^m\widehat{\boldsymbol{\beta}}_{MALR},$$

where $\boldsymbol{X}_i^m$ represents the covariates matrix corresponding to the missing entry.

This imputation approach enhances the robustness of the estimates by integrating information from both tail-sensitive MLAD and efficiency-focused GLS estimators. As a result, the imputed values are less likely to be distorted by outliers or non-normality, and are better suited for downstream statistical analysis, especially in longitudinal studies.

5. SIMULATION STUDY

The main aim of this simulation is to evaluate the performance of the proposed method. Also, the proposed method is compared with other imputation methods such as mean, median, the last observation carried forward (LOCF), the K-Nearest Neighbor (KNN), regression imputation (RM), and the Hot Deck method (HOT). The comparison covers heavy-tailed and skewed distributions. The multivariate t-distribution is chosen as an example of heavy-tailed distributions, with different levels of degrees of freedom. The chi-square distribution,with different levels of degrees of freedom, is an example for skewed distributions.

5.1. Simulation Setting and Data Generation

Three sample sizes were assumed to generate data; $n_1 = 30$, $n_2 = 100$, and $n_3 = 150$. with five time points, and subjects were divided randomly into two groups.The data is simulated from the marginal model

$$Y_i = X_i\beta + \varepsilon_i, \text{ for } i = 1, \dots, n$$

where Y_i is a vector of responses of dimension 5×1 and β is a vector of parameters of dimension 6×1. The parameters are fixed at

$$\beta = \begin{bmatrix} 4 \\ 3 \\ 2 \\ 0.5 \\ 0.8 \\ 5 \end{bmatrix}$$

The matrix$X_i$is of dimension 5×6. *Time* and *Group* are used as covariates. *Time*takes the values 1, 2, 3, 4, and *Group*takes the values 0 (first group) and 1 (second group). The variance-covariance matrix is left unstructured and fixed at

$$V_i = \begin{bmatrix} 1 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.3 & 1 \end{bmatrix}$$

The parameters (β and V_i) were adopted from [3], who used similar settings in their simulation study. This choice helps evaluate the performance of the proposed method under conditions consistent with previous literature.

The $\varepsilon_i$is generated from the following distributions:

1. The multivariate t-distribution with mean 0 and scale matrix $V_i * \frac{df-2}{df}$. The degrees of freedom are assumed to be 4, 20, and 40.
2. The Chi square distribution with degrees of freedom of 4, 20, and 40.

The missing data pattern is dropout and the missing data mechanism is missing at random (MAR). Additionally, two missing data percentages of 20% and 40% are considered. In order to create MAR missingness for the longitudinal outcome the linear logistic regression model is used, that is

$$\text{logit } p(R_{ij} = 1/\psi) = \psi_0 + \psi_1 y_{ij-1}$$

whereψ parameters of missingness indicators, and the vector of $\boldsymbol{\psi}_y$=(ψ_0,ψ_1) take two different values: $\boldsymbol{\psi}_y = (-3, 0.1)$, for low missingness rate,and $\boldsymbol{\psi}_y = (-3, 0.3)$, for high missingness rate.

5.2. Simulation Results

The performance of methods is compared using three evaluation measures. They are the mean absolute percentage error (MAPE), the normalized root mean squared error (NRMSE), and the mean absolute error (MAE). The mean absolute error (MAE) measures the average magnitude of errors between imputed and actual values, providing a straightforward interpretation of prediction accuracy in the same unit as the data. It is less sensitive to large errors compared to squared error measures, making it suitable for non-normally distributed data. It is calculated as in [13],

$$MAE = \frac{1}{r}\sum_{i=1}^{r}\left|y_i^{act} - y_i^{imp}\right|.$$

The median absolute percentage error (MAPE) expresses the prediction error as a percentage, allowing for comparison across datasets with different scales. It is widely used for evaluating imputation and forecasting accuracy because it provides an intuitive measure of relative performance. It is calculated as in [12],

$$MAPE = 100 * \frac{1}{r}\sum_{i=1}^{r}\left|\frac{y_i^{imp} - y_i^{act}}{y_i^{act}}\right|.$$

The normalized root mean squared error (NRMSE)normalizes the root mean squared error by the range or standard deviation of actual values, enabling fair comparison across datasets with varying magnitudes. It captures both the variance and magnitude of errors, making it a robust indicator of imputation quality. It is calculated as in [9],

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{r}\left(y_i^{imp} - y_i^{act}\right)^2}{r}} * \sigma^{-1},$$

where y_i^{imp} and y_i^{act} are imputed and actual measurement values, respectively, for $i = 1, \dots, r$, r is the total number of imputed values, and σ is the standerd error for actual data [9].

Tables **1** and Figure **1** present the results for sample size of 30 and missingness rate percentage of 20% for multivariate t-distribution with different degrees of freedom. The results show that the proposed method has the lowest value of the three evaluation measures (MAPE, NRNSE, and MAE) if the degrees of freedom is 4 and 20. When the degrees of freedom increased to 40 the proposed method still has the best performance in terms of MAPE, however, it has similar performance to RM in terms of the NRNSE and MAE.

Tables **2** and Figure **1** present the results for sample size of 30 and missingness rate percentage of

Table 1: Performance Comparison in Terms of MAPE, NRMSE and MAE for Multivariate T-Distribution with Sample Size N= 30 And 20% Missing Percentage

		df=4			df =20			df =40		
Evaluation measure		**MAPE**	**NRMSE**	**MAE**	**MAPE**	**NRMSE**	**MAE**	**MAPE**	**NRMSE**	**MAE**
Imputation methods	MALR	10.206	0.021	1.081	7.835	0.008	0.872	13.235	0.093	1.034
	RM	12.924	0.153	1.285	8.23	0.03	0.894	13.423	0.072	1.031
	Median	25.706	0.196	2.296	22.529	0.183	2.308	23.601	0.495	2.142
	Mean	22.48	0.114	2.261	23.674	0.224	2.35	25.435	0.335	1.947
	HOT	51.663	0.073	3.116	25.158	0.513	2.647	21.096	0.196	2.325
	LOCF	27.34	0.885	2.571	19.02	0.504	1.693	31.279	1.107	2.713
	KNN	86.442	2.161	6.28	89.439	2.64	7.164	87.88	3.212	7.196

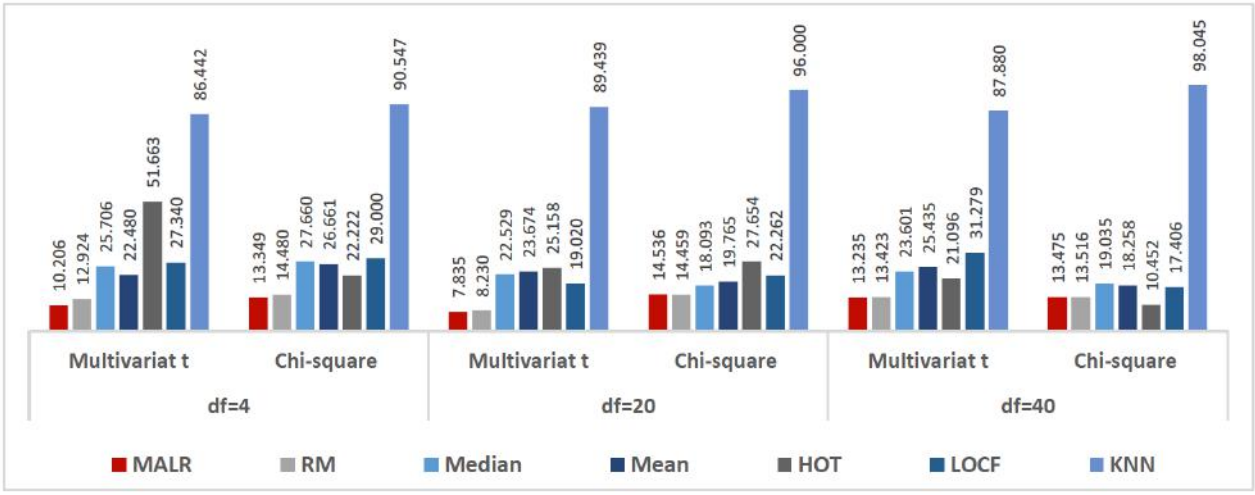

Figure 1: Performance comparison in terms of MAPE for multivariate t and chi-square distribution with sample size n= 30 and 20% missing percentage.

20% for Chi-square distribution with different degrees of freedom. The proposed approach has the best performance in terms of MAPE, NRNSE, and MAE values if the degrees of freedom is 4. The proposed approach still has the best performance in terms of NRNSE and MAE when the degrees of freedom increased to 20. However, the proposed approach has a similar performance to the RM method in terms of MAPE in this case. When the degrees of freedom increased to 40, the proposed approach has the best performance in terms of MAE. The proposed approach has a similar performance to the RM method in terms of MAPE and NRNSE.

Table **3** and Figure **2** present the results for the multivariate t-distribution at the sample size of 100 and missing percentage equal to 20%. The results show that the proposed method has the best performance in terms of the three evaluation measures. In the same time the regression method has the same performance.

Table **4** and Figure **2** present the results of the Chi-square distribution at sample size of 100 and missingness rate of 20%. It is clear from the results that the proposed approach has the best performance among all other methods, in terms of all the evaluation measures.

Tables **5** and Figure **3** present the results for multivariate t-distribution for a sample size of 100 and missingness rate of 20%. The results show that the proposed method MALR, in addition to the RM method, has the best performance in terms of all evaluation measures. The RM method has the same performance. The results in Table **6** and Figure **3** show that the proposed MALR method has the best performance among all other methods. The RM method has the same performance.

As can be seen from the results above, when the missing percentage is low, such as 20%, the proposed MALR approach performs well for small sample sizes. Although the MALR method still produces good results as the sample size increases, the RM approach yields results that are nearly identical to those of the MALR method. The simulation results for high missingness rate (40%) are very similar to those of low missingness rate (20%). However, the results are deleted for parsimony and smaller number of tables.

Table 2: Performance Comparison in Terms of MAPE, NRMSE and MAE for Chi-Square Distribution with Sample Size N= 30 And 20% Missing Percentage

		df=4			df =20			df =40		
Evaluation measure		MAPE	NRMSE	MAE	MAPE	NRMSE	MAE	MAPE	NRMSE	MAE
Imputation methods	MALR	13.349	0.180	1.785	14.536	0.147	4.095	13.475	0.186	8.639
	RM	14.480	0.229	1.873	14.459	0.264	4.164	13.516	0.180	8.643
	Median	27.660	0.315	3.089	18.093	0.172	4.982	19.035	0.208	9.643
	Mean	26.661	0.456	3.230	19.765	0.320	5.064	18.258	0.164	9.564
	HOT	22.222	0.218	2.642	27.654	0.227	7.998	10.452	0.406	8.696
	LOCF	29.000	0.658	3.219	22.262	0.813	8.232	17.406	0.428	9.479
	KNN	90.547	2.657	9.429	96.000	4.302	24.632	98.045	4.279	46.847

Table 3: Performance Comparison in Terms Of MAPE, NRMSE and MAE for Multivariate T-Distribution with Sample Size N= 100 And 20% Missing Percentage

		df=4			df =20			df =40		
Evaluation measure		MAPE	NRMSE	MAE	MAPE	NRMSE	MAE	MAPE	NRMSE	MAE
Imputation methods	MALR	17.396	0.227	1.191	16.541	0.230	0.986	15.900	0.058	0.830
	RM	17.005	0.241	1.198	16.499	0.231	0.988	15.893	0.052	0.831
	Median	23.593	0.730	1.560	19.695	0.229	1.091	20.619	0.600	1.150
	Mean	24.073	0.749	1.576	19.049	0.312	1.126	20.810	0.607	1.154
	HOT	38.168	0.664	2.084	24.251	0.080	1.470	24.976	0.235	1.475
	LOCF	49.735	1.294	2.580	38.326	1.328	2.070	42.795	1.623	2.043
	KNN	77.866	1.996	3.344	79.776	2.804	3.873	78.254	3.094	3.683

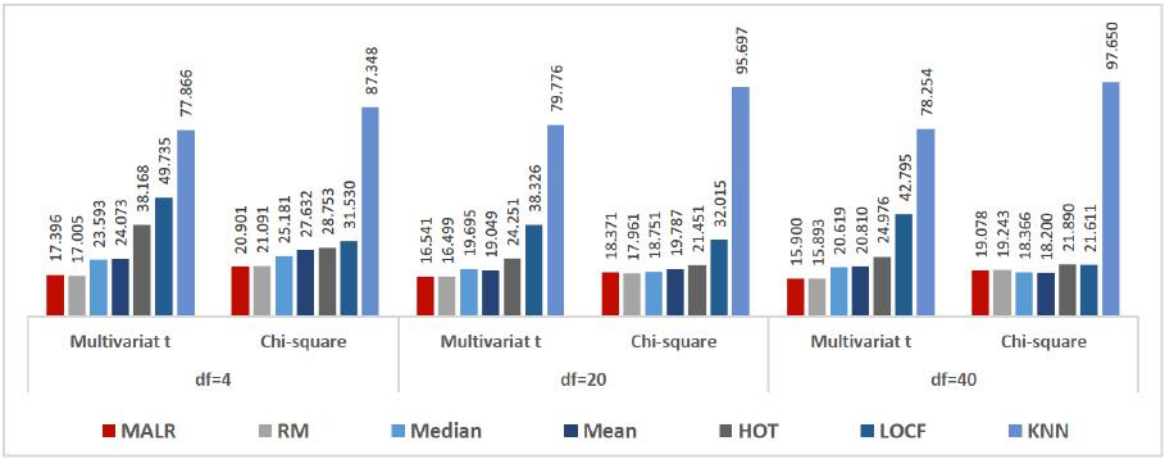

Figure 2: Performance comparison in terms of MAPEfor multivariate t and chi-square distribution with sample size n= 100 and 20% missing percentage.

Table 4: Performance comparison in terms of MAPE, NRMSE and MAE chi-square distribution with sample size n= 30 and 20% missing percentage

		DF=4			DF =20			DF =40		
Evaluation measure		MAPE	NRMSE	MAE	MAPE	NRMSE	MAE	MAPE	NRMSE	MAE
Imputation methods	MALR	20.901	0.106	1.875	18.371	0.113	5.409	19.078	0.055	8.144
	RM	21.091	0.197	1.929	17.961	0.126	5.421	19.243	0.056	8.148
	Median	25.181	0.211	2.118	18.751	0.201	5.538	18.366	0.135	8.193
	Mean	27.632	0.450	2.349	19.787	0.276	5.676	18.200	0.194	8.328
	HOT	28.753	0.332	2.871	21.451	0.210	7.198	21.890	0.232	11.874
	LOCF	31.530	0.687	2.857	32.015	0.491	8.000	21.611	0.205	11.651
	KNN	87.348	2.900	7.358	95.697	3.598	23.390	97.650	4.272	42.206

Table 5: Performance Comparison in Terms Of MAPE, NRMSE and MAE for Multivariate T-Distribution with Sample Size N= 150 And 20% Missing Percentage

		df=4			df =20			df =40		
Evaluation measure		**MAPE**	**NRMSE**	**MAE**	**MAPE**	**NRMSE**	**MAE**	**MAPE**	**NRMSE**	**MAE**
Imputation methods	MALR	23.128	0.013	1.404	15.579	0.001	0.824	14.419	0.135	0.813
	RM	22.926	0.055	1.410	15.663	0.002	0.825	14.433	0.136	0.813
	Median	24.033	0.387	1.639	19.017	0.613	1.161	18.836	0.408	1.077
	Mean	24.862	0.347	1.616	19.663	0.647	1.184	19.931	0.500	1.115
	HOT	39.100	0.449	2.310	25.906	0.703	1.406	27.906	0.387	1.576
	LOCF	34.896	0.752	2.026	38.320	1.385	1.940	36.358	1.462	1.956
	KNN	78.362	1.909	3.692	78.310	2.981	3.595	78.605	2.983	3.733

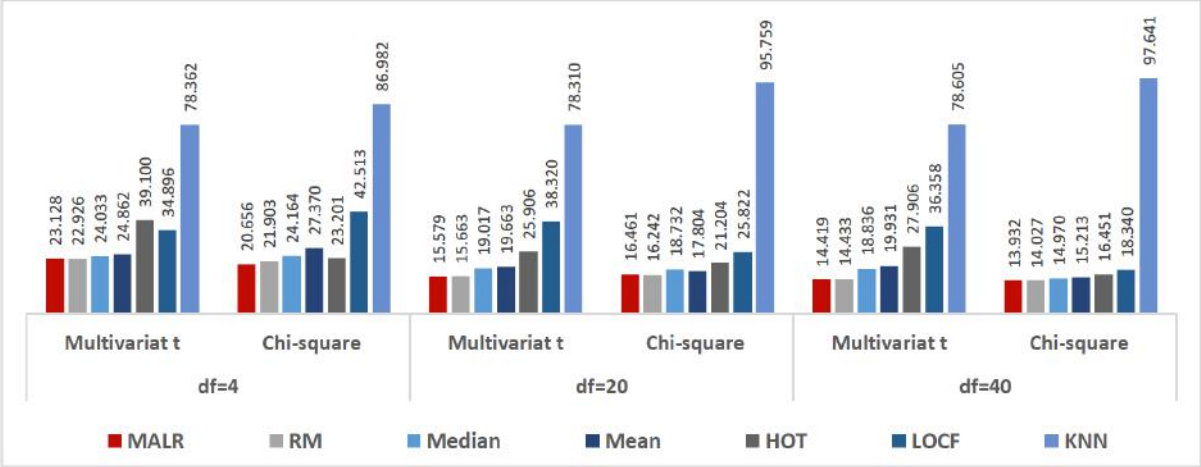

Figure 3: Performance comparison in terms of MAPE for multivariate t and chi-square distribution with sample size n= 150 and 20% missing percentage.

Table 6: Performance Comparison in Terms Of MAPE, NRMSE and MAE For Chi-Square with Sample Size N= 150 and 20% Missing Percentage

		df=4			df =20			df =40		
Evaluation measure		**MAPE**	**NRMSE**	**MAE**	**MAPE**	**NRMSE**	**MAE**	**MAPE**	**NRMSE**	**MAE**
Imputation methods	MALR	20.656	0.087	1.999	16.461	0.013	4.667	13.932	0.125	7.181
	RM	21.903	0.008	2.062	16.242	0.027	4.692	14.027	0.128	7.185
	Median	24.164	0.097	2.095	18.732	0.077	4.794	14.970	0.217	7.300
	Mean	27.370	0.324	2.333	17.804	0.187	4.960	15.213	0.242	7.351
	HOT	23.201	0.349	2.920	21.204	0.278	7.372	16.451	0.203	8.941
	LOCF	42.513	0.927	4.172	25.822	0.591	7.580	18.340	0.263	9.231
	KNN	86.982	2.707	7.422	95.759	3.808	23.314	97.641	4.781	42.221

6. APPLICATION: LABOR PAIN DATA

The proposed robust imputation method is applied to labour pain data. These data have been discussed in [2, 3, 4]. The purpose of this dataset is to evaluate the impact of two treatments on reducing labour pain for mothers. 83 labouring women were randomized to receive either a placebo (40 women) or a new pain medicine (43 women). The treatment is administered when the cervical dilatation during labour reached 8 cm. The pain was self-reported, at 30-minute intervals for a total of 3 hours (6 time-points), on a scale of 0 to 100, where 0 means “no pain”. There are missing data points in this dataset, particularly in the latter measurements. [2] and [4] demonstrate that the full dataset is skewed to the right. The full data can be found in Appendix III of [2].

Pain intensity was self-reported every 30 minutes for a total of 3 hours (resulting in six time points per participant) on a continuous scale from 0 to 100, where 0 indicates no pain and 100 indicates extreme pain.

The repeated measures nature of this dataset makes it a suitable candidate for longitudinal analysis. The dataset exhibits a monotone (dropout) missing pattern, particularly in later time points, which aligns with the Missing at Random (MAR) mechanism assumed in our analysis. In this article, we used the first three time points, where the data were mostly complete, to provide a balanced comparison across imputation methods. The variables used in the analysis were:

1. *Pain*: the continuous response variable indicating the level of labor pain,
2. *Time*: a numeric variable representing the time point of measurement (1 to 3),
3. *Treatment*: a binary variable coded 0 for placebo and 1 for the new medication, and
4. *Subject ID*: a unique identifier for each woman to account for within-subject correlation.

This dataset is known to exhibit right-skewed and heavy-tailed behavior, making it ideal for testing the robustness of the proposed MALR imputation method. Previous analyses (e.g. [2, 4]) have shown non-normal residuals, and the kurtosis of errors in our analysis further confirmed this characteristic. These features provide a realistic setting for assessing the performance of imputation methods under non-normal conditions.

Finally, the dataset consists of 30 women in the placebo group and 35 women in the treatment group. To apply the proposed technique, we dropped some observations from the data. It is assumed that the missing data pattern is dropout and the missing data mechanism is random (MAR). This is conducted using R package. We use the following model to analyze the data, in accordance with [3, 4],

$$y_{ij} = \beta_0 + \beta_1 t + \beta_2 x_i + \beta_3 t \times x_i, \text{ for } i = 1, \ldots, 65, j = 1,2,3 ,$$

where y_{ij} is the amount of pain for patient i at time-point j, x_i is the treatment variable with 0 for placebo, and 1 for treatment, and$t$ is the time which ranges from 1 to 3.

Figure **4** shows that the distribution of the pain for placebo and treatment group is positively skewed. The average kurtosis of the errors resulting from the GLS estimator and the MLAD estimator is 22.9. This means that the distribution of errors seems to be heavier than normal. Hence, the MALR estimator is expected to perform better in this case.

We applied the proposed MALR imputation approach to handle missing values.The performance of the method is evaluated by comparing the results of the observed data to the imputed data obtained through other imputation methods. In simulation studies, the evaluation measures that were used are median

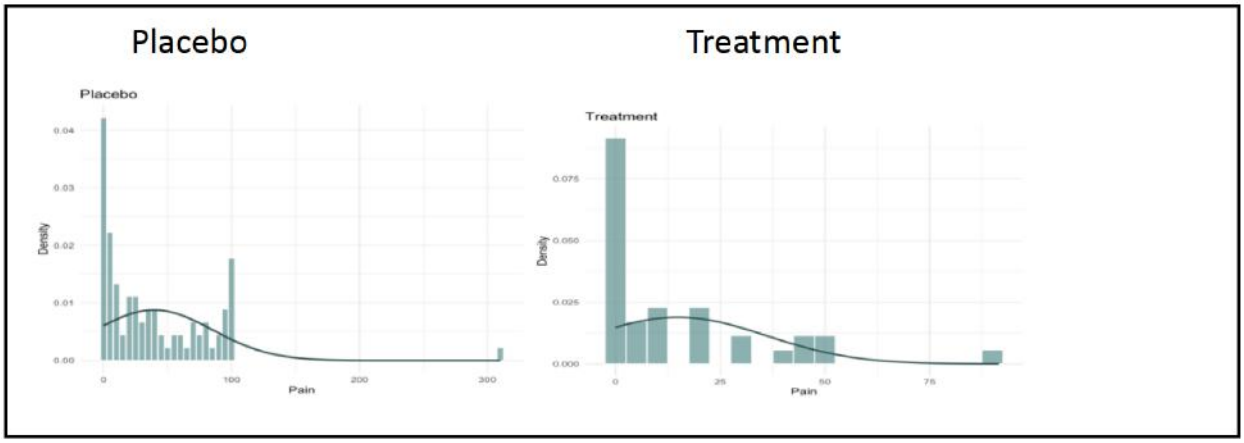

Figure 4: Histogram of responses in the placebo and treatment group.

Table 7: Performance Comparison in Terms of NRMSE and MAE for Each Imputation Method for Labour Pain Data

Evaluation measure		NRMSE	MAE
Imputation methods	MALR	0.115	14.233
	RM	0.366	14.812
	Median	0.746	20.100
	Mean	0.014	17.200
	HOT	0.911	26.600
	LOCF	0.637	16.250
	KNN	0.307	11.425

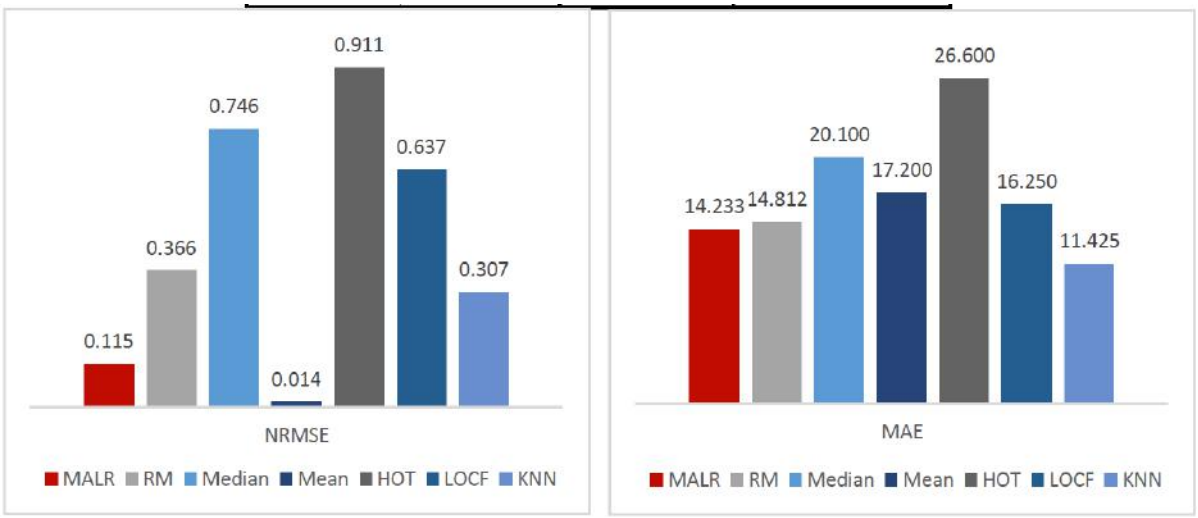

Figure 5: Performance comparison in terms of NRMSE and MAE for each imputation method for labour pain data.

absolute percentage error (MAPE), normalized root mean squared error (NRMSE) and mean absolute error (MAE). However, in application study, MAPE was not used because it divides by the actual values, some of which are equal zero. So, in this application study, the evaluation measures that were used are NRMSE and MAE.

Table **7** and Figure **5** present the results of NRMSE and MAE for each imputation method for labour pain data. The two measures of performance for the proposed technique, the MALR method - NRNSE and MAE, are lower than those for other methods. The NRMSE measure for the proposed MALR imputation method and the mean imputation method yield nearly identical results. But the MAE for the KNN imputation method is less than the MAE for the proposed MALR imputation method. In summary, the results show that the proposed imputation method performs well comparable to its competitors.

5. CONCLUSION

Several challenges appear when analyzing longitudinal data. One of these challenges is the presence of missing values, which leads to loss of information, biased estimates, and misleading results. Imputation methods can be used to handle missing values. In practice, normality assumption of the response may be violated. So, there is a need for an imputation method that can cope with non-normal responses and outliers.

In this article we propose a robust imputation method that can be used in the presence of outliers, heavy-tailed distributions, or skewed distributions. In literature there is no single robust imputation technique has been suggested in longitudinal context. Hence, this method can be considered the first single robust imputation method in longitudinal data context.

The proposed imputation method is quite simple method and can be used in different areas where missing values are present. For example, in health records usually we encounter missing records where imputation is needed. In health surveys where some participants refuse to answer some questions which results in missing values. The proposed imputation method can be used in these circumstances.

Despite its promising performance, the suggested MALR-based imputation method has a few drawbacks that should be noted. First, the current framework focuses on single imputation under the Missing at Random (MAR) assumption and dropout (monotone) missingness pattern. Although these settings are common in longitudinal studies, real-world data may exhibit more complex non-monotone (intermittent) or Missing Not at Random (MNAR) mechanisms, which are not explicitly addressed here. In addition, the method has not yet been extended to a multiple imputation framework, which would allow for proper propagation of imputation uncertainty in inferential analyses. Future research could therefore aim to (i) generalize the MALR estimator to multiple imputation schemes, (ii) adapt it for mixed missingness mechanisms, and (iii) explore its integration with Bayesian or machine learning–based approaches for higher-dimensional longitudinal data.

DECLARATION STATEMENTS

As the data is open source, there are no experiments on humans conducted by the authors.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

CONSENT TO PUBLISH

Authors affirm there is no figure of any participant in the article.

DATA AVAILABILITY

The dataset is shared open source. The data are available in Appendix III of Davis C. Semi-parametric and non-parametric methods for the analysis of repeated measurements with missing values. Statistics in Medicine 1991; 10 (12): 1959-1980.

REFERENCES

[1] Carpenter JR, Kenward MG. Multiple imputation and its application. Wiley, UK 2013. https://doi.org/10.1002/9781119942283

[2] Davis C. Semi-parametric and non-parametric methods for the analysis of repeated measurements with missing values. Statistics in Medicine 1991; 10(12): 1959-1980. https://doi.org/10.1002/sim.4780101210

[3] Gad AM, Ibrahim WIM. An adaptive linear regression approach for modeling heavy-tailed longitudinal data, Communications in Statistics - Simulation and Computation 2020; 49(5): 1181-1197. https://doi.org/10.1080/03610918.2018.1491990

[4] He X, Fu B, Fung WK. Median regression for longitudinal data. Statistics in Medicine 2003; 22(23): 3655-3669. https://doi.org/10.1002/sim.1581

[5] He Y, Zhang G, Hsu CH. Multiple imputation of missing data in practice. CRC Press, UK 2022. https://doi.org/10.1201/9780429156397

[6] Jadhav A, Pramod D, Ramanathan K. Comparison of performance of data imputation methods for numeric dataset. Applied Artificial Intelligence 2019; 33(10): 913-933. https://doi.org/10.1080/08839514.2019.1637138

[7] Jahangiri M, Kazemnejad A, Goldfeld KS, *et al.* A wide range of missing imputation approaches in longitudinal data: A simulation study and real data analysis. BMC Medical Research Methodology 2023; 23: 161. https://doi.org/10.1186/s12874-023-01968-8

[8] Li J, Guo S, Ma R, *et al.* Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. BMC Medical Research Methodology 2024; 24: 41. https://doi.org/10.1186/s12874-024-02173-x

[9] Li J, Yan XS, Chaudhary D, *et al.* Imputation of missing values for electronic health record laboratory data. NPJ Digital Medicine 2021; 4: 147. https://doi.org/10.1038/s41746-021-00518-0

[10] Rubin DB. Inference and missing data. Biometrika 1976; 63(3): 581-592. https://doi.org/10.1093/biomet/63.3.581

[11] Salgado CM, Azevedo C, Proença H, Vieira SM. Missing data. In: secondary analysis of electronic health records. Springer, Cham 2016; 143-162. https://doi.org/10.1007/978-3-319-43742-2_13

[12] Simpson L, Wilson T, Shalley F. The Shelf Life of Official Sub-National Population Forecasts in England. Applied Spatial Analysis and Policy, 2020; 13(1): 715-737. https://doi.org/10.1007/s12061-019-09325-3

[13] Twumasi-Ankrah S, Odoi B, Pels WA, Gyamfi EH. Efficiency of imputation techniques in univariate time series. International Journal of Science, Environment and Technology 2019; 8(3): 430-453.

Received on 16-10-2025 Accepted on 18-11-2025 Published on 08-12-2025

https://doi.org/10.6000/1929-6029.2025.14.70