Missing Data in Regression Analysis: A Review

Shreen Ali Hussein
shreen.a@uobaghdad.edu.iq
Saad Kadem Hamza
saad.hamza@coadec.uobaghdad.edu.iq
University of Baghdad
Corresponding Author : Shreen Ali Hussein

Abstract : The problem of missing data is a major obstacle for researchers in the process of data analysis in various fields, and this problem appears frequently in all fields of social, medical, astronomical studies, clinical trials, and others. The presence of such a problem within the data to be studied will negatively affect its analysis and then lead to misleading conclusions, and these conclusions result from the great bias caused by this problem. Therefore, this work provides a comprehensive analysis of the different methods used to solve the problem of missing data in databases. It identifies the different types of missing data and points out the most common types of regression analysis. It also aims to introduce the reader to many methods for solving the problem of missing data in regression analysis, while explaining how these methods affect the final conclusions of the study.
Paper type: Promotion paper
Keywords: missing data, The Regression Model's, MCAR,MAR,MNAR, Handling Missing Data

Introduction:

In research studies and surveys, missing data is inevitable and often leads to inaccurate conclusions. A variety of factors may result in missing data, giving rise to distinct kinds of missing data, such as MCAR, MAR, and MNAR. Regardless of the cause of the missing data, it must be disregarded during the analysis phase since this might skew the outcomes [1].

A number of protocols have been developed to handle missing data in databases. The most straightforward way is to exclude all observations, even those with missing data (complete-case method); however, this is wasteful and may result in biased findings. Furthermore, it cannot be used if a significant amount of data is absent, since doing so would result in information waste. Imputation, weighing processes, and available-case procedures are other methods. In the MCAR and MAR scenarios, the missing data is replaced with projected values via the use of imputation techniques [46]. In the process of imputation, there are two types of procedures: model-based techniques and non-model-based techniques [2]. [35] Mean, median, mode and hot-deck imputation are examples of non-model-based methods [1], [2].

These methods may cause bias in the findings of statistical processes by lowering the variance estimations. Conversely, MBD [45] includes ANN approaches, multiple imputations, regression-based methods, and expectation maximisation [2], [47].

The actual effects of missing data on regression were investigated in [3] KEEL and UCI datasets (Abalone, Arfoil, Bike, California, Compactiv, Mortage, Wankara, and Wine) were used to analyse its effects. To impute missing data, a variety of techniques were used, including decision trees, random forests, adaboosts, KNN, SVM, and NN[43],. A simulated investigation of the various datasets led to the conclusion that missing data may have a big impact. The study's findings show that, when it comes to the regression of data with missing values, the K-NN method performs better than others [3].

Since the Random Forests approach performs better in terms of prediction accuracy and computing efficiency, it was chosen to evaluate the effect of missing data. Three techniques were used to analyse the impact: testing classifier performance, analysing statistical differences between imputed and genuine values [49], and using logistic regression for probability prediction. Data sets were constructed with varied quantities of imputed variables. The results show that, with the right imputation strategies, the algorithm is still resistant to missing data. [2].

This paper covers regression model definitions and common types, reasons for missing data, ignorable and nonignorable missingness, missing data patterns, and eight methods for handling missing data: Complete-case and available-case analysis, single and multiple imputation, maximum likelihood, Bayesian methods, KNN, regression trees, and random forests [48]. It emphasises how different techniques for managing missing data impact regression modelling outcomes. The aim is to familiarise readers with various approaches to addressing missing data in regression analysis and illustrate their effects on results [38].

2.The Regression Model's Definition

Regression models assess the impact of covariates on the dependent variable, which is determined by the types of variables (continuous, binary, categorical, or counts) and the response variable (continuous, binary, or categorical). These models are characterised by a probabilistic connection with random errors between the variables. [4]. [44] .

2.1 Types of Regression Models

Different regression models can be summarized as follows:

1- Linear Models

Data

(y_i, x_i), i = 1, …. , n, with continuous variables y and x.

Model

$y_i = \beta_o + \beta_1 x_i + \varepsilon_i$, i = 1, …. , n. .....(1)

The errors $\varepsilon_1, …., \varepsilon_n$ are independent and identically distributed with $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ [4].

2- Logistic Regression Model for Binary Response Variables

Data

$(y_i , x_{i1}, …. , x_{ik})$, where i = 1, …. , n

y is a binary response variable ϵ {0, 1} for continuous or coded covariates $x_1, …. , x_k$) [4].

Model

Probability $P(y_i = 1) = \exp(\lambda_i) / (1 + \exp(\lambda_i))$ .....(2)

Where λ_i is a linear predictor given by

$$\lambda_i = \beta_o + \beta_1 x_{i1 + … +} \beta_k x_{ik} \quad .....(3)$$

3- Multiple Linear Model

Data

(y_i, x_{ki}), k = 1, 2, …. , p, with continuous variables y and x.

Model

$y_i = B_0 + B_1x_{1i} + B_2x_{2i} + B_3x_{3i}+…….+ B_px_{pi} + e_i$

Where:

B_0 is the constant term and B_1 to B_p are the coefficients relating the *p* explanatory variables to the response variable [5].

2.2 Reasons and Types of Missing Data

Understanding missing data is crucial for effective data management. Factors such as survey responses, data loss, improper recording, or intentional design, including planned missingness, can influence it and potentially bias the analysis. [8]. Different types of missing data influence its impact on final conclusions. Common classifications include:

1- MCAR

The recorded data and the missing values do not vary in any systematic way [1]. Put differently, the causes contributing to a missing data item happen entirely at random and are independent of both observable and unobservable variables. [6].

2- MAR

Differences in the other variables in the data set may account for the systematic disparity between the observed and missing data [1]. Put another way, variables with complete information in the data set may completely account for missing data that happened. The missingness in this instance is not coincidental [6].

3- MNAR

For reasons that are unknown to the researcher, the probability of a data point being absent fluctuates [9]. Stated otherwise, the cause for missingness is connected to the missing value [6].

2.3 Ignorable and Nonignorable

Nonignorable missing data requires a model for missingness, while ignorable missing data requires at least MAR, and improving estimators for parameters in the data model is not necessary. To identify missing values in datasets, identify variables with missing data, quantify their extent, and consider missingness types. Accurate assumption of missingness types affects result reliability. Calculate missing value percentages, exclude less critical variables, and determine missing value locations. . [11]. [30]

2.4. Missing Data Patterns

Some missing data handling techniques are applicable universally, while others are restricted to specific patterns. Figure 1, illustrates four instances of missing data patterns among variables (X's). Univariate missing data, where only one X (e.g., X_1) has missing values, is depicted. This aligns with the general case of monotone missing data shown in Figure 2, where columns may be organised such that for each observed instance of X_j (where j = 1,..., p), X_{j+1} is observed.

Figure 3, illustrates a pattern where two Xs (X_1 and X_2) are never observed together. When combining two samples with information on (X_1, Y) and (X_2, Y) into one database, this pattern emerges. Estimating regression from this pattern requires an assumption about the conditional association of X_1 and X_2, given X_3 and Y.

Lastly, Figure 4 represents a broad pattern without any distinct structure. [12].

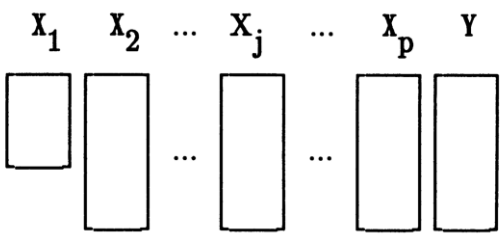


Fig. (1) Univariate missing data pattern

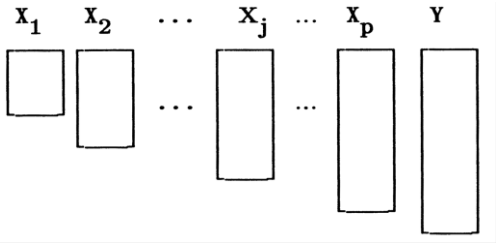


Fig. (2) Monotone missing data pattern

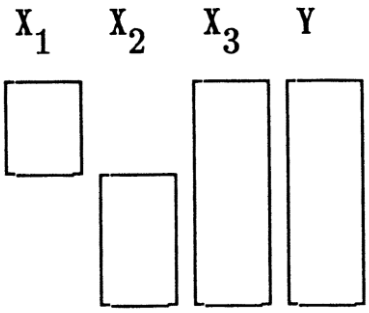


Fig. (3) Special missing data pattern with unidentified parameters

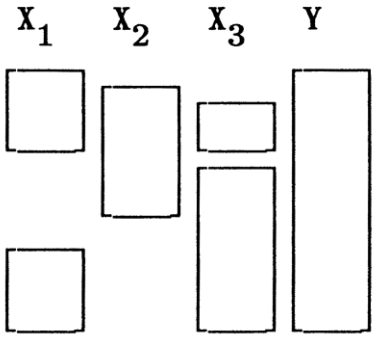


Fig. (4) Missing data general pattern

2.5 Methods of Handling Missing Data

In order to determine the best techniques for handling missing data, take into account evaluation criteria and adhere to guidelines like minimizing bias, making the most use of the information at hand, and producing precise estimates of uncertainty for statistical analysis, like p-values, standard errors, and confidence intervals. [10].

1- CCA

Statistical packages like linear regression use listwise deletion to model datasets, removing missing variables and leaving only complete cases, assuming missing data are MCAR. [7].

1.1 Effect of applying Complete-Case Analysis method on outcome results

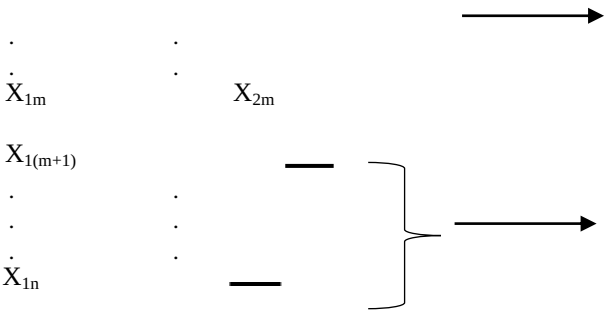
MCAR data is more probable to be applied when a data set has few missing values; which means that when only few cases are missing, there is more likelihood that the available complete cases represent the population [7]. In instances where there is significant missing data, possibly due to multiple covariates with incomplete observations, a few intact cases may remain for analysis. However, relying solely on complete-case analysis may not be appropriate in such scenarios. [7]. The complete-case analysis method offers ease of implementation but cannot guarantee sufficient data for analysis. [7].

2- ACA

In pairwise deletion, all accessible data are employed to estimate model parameters. For instance, in univariate descriptive statistics with missing data, means and variances are computed for observed variables across the dataset (utilizing available case analysis). When examining bivariate or multivariate relationships through pairwise deletion, all cases contribute to estimating the mean of X_1, while only complete cases are considered for estimating X_2 and the correlation between X_1 and X_2. (Fig. (5) The text presents a two-variable dataset with one missing variable, utilising different case sets to estimate the parameters of interest. [7].

X_{11}	X_{21}
X_{12}	X_{22}
.	.

m complete cases



Available case estimates:

$\bar{\chi}_1 = \frac{1}{n} * \sum_{i=1}^{n} x_{1i}$

$\bar{\chi}_2 = \frac{1}{m} * \sum_{i=1}^{m} x_{2i}$ .....(6)

$s_1^2 = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{\chi}_1)^2}{n-1}$ .....(5)

$s_2^2 = \frac{\sum_{i=1}^{m} (x_{2i} - \bar{\chi}_2)^2}{m-1}$ .....(6)

$$r_{xy}^2 = \frac{1}{m-1} \frac{\sum_{i=1}^{m} (x_{1i} - \bar{x}_{1(m)})(x_{2i} - \chi_2)}{s_{1(m)}\, s_2} \quad \text{.....(7)}$$

where $\bar{x}_{1(m)}$ and $s_{1(m)}$ -are the mean and the standard deviation of x_1 calculated from the m complete cases.

2.1 Impact of applying Available Case Analysis method on outcome results

Available case analysis uses MCAR data to generate unbiassed parameter estimates but may result in biased estimates if MAR data is used. It may be more efficient than full case analysis, but simulation studies suggest a decrease in efficiency. Accurate standard error estimates are difficult to obtain due to sample sizes and missing data patterns. [10].

3-Imputation

Imputation replaces missing values in data sets with estimates, with single and multiple types available. Single imputation imputes one value for each missing item, allowing additional uncertainty. [1].

3.1 Single Imputation and its effects on the estimated parameters

The method of substituting missing values by their mean value is a common one, but it is known to produce biassed estimates. On the other hand, the linear regression model is better for imputed missing values but has some issues, such as biased estimates of parameters that depend on variances, such as regression coefficients. [8], [10].

Linear regression models estimate the relationship between a data point and its associated variable using the least-squares method. However, regression imputation has a drawback as it infers the most probable value of missing data without providing information about the uncertainty of the imputed value. [13].

3.2 Stochastic regression imputation

Is a method that replaces missing values in a dataset with predicted values from regression analysis, using complete cases and a random residual term. When choosing the appropriate model, it outperforms traditional regression imputation. [8], [13]. To simulate a missing value, HDI substitutes observed values from a randomly chosen example. This approach, however, may overstate standard errors and skew variable connections, inflating test statistics and lowering p-values. This intrinsic uncertainty is not taken into account by conventional statistical software, which exacerbates the issue as missing data fractions rise. As such, care should be taken while using traditional imputation techniques. [8], [10].

3.3 MI and its impact on the estimated parameters

This approach involves sampling multiple values from a predictive distribution and repeating complete-data analyses I times using one of the imputed substitutes, instead of relying on a single mean for each missing value. [11]. MI faces challenges such as thousands of iterations for three to five completed data sets, requiring a minimum of five sets for unbiased estimates, and the need for specialized software to handle the costly computing time.
[7].

4- ML Method Using the EM Algorithm

This method regards all variables with missing values as random variables within a defined model structure. Its primary strength lies in yielding estimates with reduced standard errors. [8], [14].

ML models specify the expected values, variances, covariances, and probability distributions of dependent variables and covariates with missing values. We can obtain robust estimates and standard errors by using the scoring function as an estimating equation and calculating standard errors via replication. [14]. [36] EM is a technique that maximises the likelihood of predicting missing values in a data model, especially when known, allowing unbiased predictions and simplifying parameter estimation. It involves estimating parameters, missing values, and using the filled-in dataset, ending when stable estimates are reached. [15]. [31]

4.1 Impact of applying Maximum Likelihood method on outcome results

The EM algorithm offers the benefit of yielding unbiased, or close to unbiased, estimates for means, variances, and covariances. Additionally, it performs effectively even if the assumption of a multivariate normal distribution of observations is incorrect. [15].

5- BAYESIAN METHODS

Linear regression and ML methods are insufficient for inference with small samples. A proposed solution involves incorporating a prior into the likelihood and basing inference on the posterior distribution. The Bayesian approach is used for multivariate problems with missing dependent variables, but its applicability to regression models with missing covariates is limited. [12], [37],

6- KNN

NN algorithms are good at filling in missing data by using values from similar cases in the whole dataset to replace missing values. This makes the values more believable and closer to the truth. [16]. [38].

6.1 How to implement the *K-NN* procedure

To proceed, two initial tasks must be completed: defining the computation method and distance measures for NN, and establishing the procedure for obtaining an imputed value using NN.

6.2 Distances and the NN calculation

1- Let p be the number of variables and n be the number of observations. X = (x_{is}) yields the matching n × p data matrix, where x_{is} represents the ith observation of the sth variable.
Let O = (ois) represent the matching n×p dummy matrix with entries.

$$o_{is} = \begin{cases} 1 & \text{if } x_{is} \text{ was observed} \\ 0 & \text{for missing value.} \end{cases}$$

The distance between two observations x_i in one row and x_j in another row in the data matrix, can be computed by using the d_q metric for the observed data, given by

$$d_q(x_i, x_j) = \left[\frac{1}{m_{ij}} \sum_{s=1}^{p} |x_{is} - x_{js}|^q \; I(o_{is} = 1) I(o_{js} = 1)\right]^{1/q} \quad \text{.....(8)}$$

where: $m_{ij} = \sum_{s=1}^{p} I(o_{is} = 1) I(o_{js} = 1)$

m_{ij} denotes the number of valid components in the computation of distances. If an is true, the value of the indicator function I(a) is 1, and if not, it is 0. The distance calculation only uses the vector components for which observations in both vectors are available.
The calculation of neighbors as follows:

$\mathbf{c_{ij}} = \{ s: I(o_{is} = 1) I(o_{js} = 1) = 1 \}$

When imputing a given value, the NN is defined by the dista .....(9) xj) [17].

1- In K-NN Regression problems, the average of the k nearest neighbours is taken to predict the missing value, where regression here is used with continuous variables [18].

6.3 Impact of using K-NN algorithms on the final findings

In terms of imputation precision and reduced errors in inferential statistics, the k-NN algorithm typically performs better than the 1-NN algorithm regardless of the framework; however, the 1-NN algorithm is the only one that can preserve the data structure, and there are data distortions for higher values of k neighbours [16].
As the dataset expands, KNN's increasing inefficiency compromises the model's overall performance [17]. Even with a considerable quantity of missing data in the training sets, the k-NN method with k = 10 produced extremely excellent results when employed for missing data imputation [19].

6.4 Advantages of K-NN algorithm [38].

- Simple Implementation
- Seamless Adaptability: The algorithm effortlessly adjusts to integrate new training samples, ensuring it remains responsive to evolving data.
- Minimal Hyperparameters: KNN only necessitates setting a k value and selecting a distance metric [18]

6.5 Disadvantages of K-NN algorithm

A lazy algorithm suffers from limited scalability, dimensionality issues, and overfitting. Optimal feature selection and dimensionality reduction can mitigate these issues. [18].

7- Regression Tree

The regression tree algorithm is a regression model used in the imputation process to predict missing values by using covariates and decisional rules. [20]. [40].
A regression tree algorithm splits data into homogeneous subsets using categorical covariates, creating a model with leaves as imputation cells for the target variable. R(t), the variance for node t, measures the impurity level of a node. It can be expressed as follows: [20]:

$$R(t) = \frac{1}{N(t)} \sum_{i \epsilon t} (y_i - \bar{y}(t))^2 \quad \text{....(10)}$$

The criterion function for split s at node t is defined by N(t), yi, the target variable value for the i-th sample, and $\bar{y}(t)$, the mean of the target variable in node t.

$\Phi(s, t) = R(t) - p_L R(t_L) - p_R R(t_R)$

The split s is chosen based on its ability to maxi ....(11) lue of Φ(s, t), which represents the impurity reduction achieved through the creation of two offspring nodes. Unless otherwise stated, we choose the split to ensure the maximum improvement of tree homogeneity. [20].

7.1 Impact of applying Decision Tree Algorithm on the outcome results

It A decision tree algorithm is a method that is resistant to outliers and missing data, but it tends to overfit data, which can be resolved by setting constraints on model parameters like tree height and pruning. [21].

8- Random Forest

The RF algorithm uses bagging and feature randomness to create an uncorrelated forest of decision trees, ensuring low correlation among them. Unlike decision trees, random forests select a subset of possible feature splits. [22]. The

RF method uses an ensemble of decision trees, each using a bootstrap sample or data sample from the training set. One-third of the sample is set aside for testing, and the dataset is given additional variability through feature randomness. The output is averaged for regression tasks, and the predicted class is determined by a majority vote for classification tasks. A forecast is produced by cross-validating the OOB sample. [22].

8.1 Effect of applying RF Algorithm on the outcome results [23], [32] , [33]

1. RF classifiers reduce overfitting risk by averaging uncorrelated decision trees, reducing overall variance and prediction error.
2. The RF classifier, due to its feature randomness, is highly effective in estimating missing values, as it maintains accuracy even when some data is missing.
3. R F simplifies evaluating feature importance using Gini importance and mean decrease in impurity (MDI), assessing the model's accuracy decrease when a variable is excluded.
4. The RF algorithm, which can handle large data sets, can be slow in processing data.
5. The RF algorithm requires more storage space due to its larger dataset processing capacity.
6. The complexity of interpreting a single decision tree's prediction is reduced compared to a forest of them [23].

9- Dependent Variable

Regarding the estimate of a regression model, there are two issues about the dependent variable that contains missing data. [10]. [34]

Do we need to include the dependent variable in the covariates used for imputing missing values in the independent variables? No, is the response. The random component in multiple imputation eliminates this bias, even though using the dependent variable to impute missing data for independent variables can lead to overestimated coefficients. Excluding the dependent variable often results in regression coefficients being attenuated towards zero. [10].

Secondly, is imputation necessary for the dependent variable itself? No, if the missing data is missing at random and no auxiliary variables are available. Imputing the dependent variable raises sampling variability, so it's preferable to remove cases with missing data on the dependent variable before imputation. However, if there are strongly correlated auxiliary variables with the dependent variable, imputing the dependent variable can enhance efficiency and mitigate bias. [10].

3. Discussion

The effect of missing data on regression modelling of datasets can be significant, as revealed by simulation studies [3]. Missing data treatment methods described in this work comprise the following procedures: (CCA, ACA, imputation (single - multiple i), maximum likelihood, Bayesian methods, K-nearest neighbours, regression trees, and random forests). Summarization of the pros and cons of each technique are given in Table (1),

Table (1) Advantages and disadvantages of various approaches to managing missing data

	Technique Name	Pros	Cons
1	CAA	Easiness of implementation	May not be viable if too missing data are present.
2	ACA	When the missing data is Missing Completely at Random, the parameter estimates are unbiased	If the missing data is MAR, the estimates of the parameters are unbiased.
3	SI	Straightforward, convenient, and swift to implement. [13]	Yields estimates with bias.
4	MI	Numerous software packages are accessible. Resilient. Primarily furnishes unbiased estimates [13]	potentially iterate through numerous cycles.
5	ML	Leads to estimates with reduced standard errors. Generates unbiased estimates for means, variances, and covariances.	Calculations can be intricate, often necessitating complex mathematical integrations. [13]
6	Bayesian methods	Capable of producing inferences with favorable frequentist characteristics	There are constraints on its use in regression models with absent variables.
7	KNN	Implementation is straightforward. Adapts effortlessly. Requires minimal hyperparameters.	Poor scalability Challenges of dimensionality Susceptible to overfitting
8	Regression trees	it remains unaffected by outliers and missing data.	It has a tendency to overfit the data. It's not a precise match for continuous data. [40]
9	RF	1-Decreased likelihood of overfitting Simplified recognition of feature significance Offers adaptability	Time-intensive procedure Demands additional resources Increased complexity

4. PERFORMANCE EVALUATION OF DATA

The imputation technique used should be as close to the true value as possible. Performance measures include minimising the NRMSE, RMSE [41]., or using Pearson's PAC. Mathematical equations describe these techniques, with better performance observed when the PAC is closer to 1. [13].

$$\text{NRMSE} = \sqrt{\frac{\sum_{n=1}^{N}(\frac{Y_o-\mu}{\sigma}-\frac{Y_i-\mu}{\sigma})^2}{N}} \quad \text{....(12)}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{n=1}^{N}(Y_o-Y_i)^2}{N}} \quad \text{....(13)}$$

$$\text{PAC} = \frac{\sum_{n=1}^{N}(Y_{i,n}-Y_{i,m})(Y_{o,n}-Y_{o,m})}{\sqrt{\sum_{n=1}^{N}(Y_{i,n}-Y_{i,m})^2(Y_{o,n}-Y_{o,m})^2}} \quad \text{....(14)}$$

Where: Y_o = observed value , Y_i = imputed value , $Y_{o,n}$ = n-th observed value ,
$Y_{i,n}$ = n-th imputed value , $Y_{o,m}$ = mean value of Y_o , $Y_{i,m}$ = mean value of Y_i

5. Additional techniques to improve the accuracy of Imputation

Multiple Imputation is a type of what is known as repeated imputation, another type of repeated imputation is called FI.

FI, particularly FHDI, reduces imputation variances compared to MI. FHDI inherits benefits from hot-deck imputation: it preserves the original data's distribution characteristics by relying on observed responses and doesn't require strict model assumptions. Its primary advantage is the ability to generate general estimates without the need to satisfy the congeniality condition required by MI. Additionally, FHDI avoids improper imputation issues by being based on the frequentist's EM algorithm. [24].

Imputation via clusterwise linear regression is a method employed to fill in missing data points. This approach involves identifying optimal clusters within the dataset and applying linear regression modelling to each cluster. By leveraging the assumption that the available dataset is representative and utilising only data points similar to the incomplete samples, imputation via clusterwise linear regression yields precise predictions for missing values. Comparative analysis against other imputation methods, such as MICE, using evaluation metrics like mean absolute error and root mean square error demonstrates that imputation via clusterwise linear regression often yields values with smaller errors, particularly in MCAR and MAR datasets with small to moderate percentages of missing values. [25].

By Using Surrogate Data[40]**,** we can handle covariates in regression analysis with nonignorable missing values. In [26], We proposed a method for estimating missing data using equations based on the conditional expectation of the outcome, taking into account both observed covariates and surrogate data values. The proposed estimator demonstrated good theoretical and empirical results. [26]. For not identifying any parametric model for the missing data mechanism,part of the observed covariates should be discarded, and extra parametric model assumptions have to be made. Those measures are worthy and not quite expensive [26]. [39]

Reference [27] The study presents a method for estimating missing data using artificial neural network clustering and L2 regularized regression with symmetric uncertainty. It improves imputation accuracy by sequentially imputing missing values, consuming less computational time. Experiments were conducted on genomic and non-genomic datasets, proving its effectiveness in high-dimensional datasets and enhancing biomedical data classification. [27]. [29]

In reference [28], A method for imputed missing traffic state data was developed using a graph aggregator-generative adversarial network. Historical road data correlation coefficients were used to create a new network with strong temporal relationships. The graph aggregate method was used to obtain spatial-temporal information. Comparative experiments showed the model's outperformance in various case studies. [28].

6. Conclusion

Data cleaning is crucial for DA to enhance efficiency and quality. Addressing missing data is essential to prevent biased predictions and incorrect outcomes. Various methods, including CCA, ACA, imputation techniques, Bayesian methods, KNN, regression trees, and random forests, have been developed to address this issue. While no single algorithm consistently outperforms others, selecting the most suitable one can significantly improve model prediction

accuracy. KNN is a powerful classification and regression method but has drawbacks like high computational costs and noise sensitivity.

7.References

1. The Effects of Missing Data on Multiple Linear & Logistic Regression, (2011). Dissertation in Statistics. The University of Leeds, School of Mathematics. [Online]. Available: https://resources.library.leeds.ac.uk/final-chapter/dissertations/maths-tpg/example6.pdf, [Accessed 4 January 2024].
2. Pantanowitz, A., & Marwala, T. (2009). Missing data imputation through the use of the random forest algorithm. In *Advances in computational intelligence* (pp. 53-62). Springer Berlin Heidelberg..
3. Marcelino, C. G., Leite, G. M., Celes, P., & Pedreira, C. E. (2022). Missing data analysis in regression. *Applied Artificial Intelligence*, *36*(1), 2032925.
4. Fahrmeirand, L., Lang, L., Kneib, T., and Marx., B., (2013). Regression: Models, Methods and Applications. (eBook), DOI 10.1007/978-3-642-34333-9
5. Tranmer, M., Murphy, J., Elliot, M., and Pampaka, M. (2020). Multiple Linear Regression, (2nd Edition);Cathie Marsh Institute Working Paper. https://hummedia.manchester.ac.uk/institutes/cmist/ archive-publications/working-papers/2020/2020-1- multiple-linear-regression.pdf.
6. WIKIPEDIA, Missing Data, [Online]. Available: https://en.wikipedia.org/wiki/Missing_data. [Accessed 5 January 2024].
7. Pigott, T. D. (2001). A review of methods for missing data. *Educational research and evaluation*, *7*(4), 353-383.
8. Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological methods*, *6*(4), 317.
9. Concepts of MCAR. MAR and MNAR. [Online]. Available: https://stefvanbuuren.name/fimd/sec-MCAR.html [accessed on 6 Jan 2024].
10. Allison, P. D., (2009). Missing Data, statistical horizons, [Online]. Available: https://statisticalhorizons.com/wp-content/uploads/2012/01/Milsap-Allison.pdf. DOI: https://doi.org/10.4135/9781412985079 [accessed on 5 Jan 2024].
11. Little, R. J. (1992). Regression with missing X's: a review. *Journal of the American statistical association*, *87*(420), 1227-1237.
12. Little, R. J. A., (2011). Regression with Missing X's: A Review, UCLA, Department of Statistics Papers, [Online]. Available: https://escholarship.org/uc/item/84j7c2w5 [accessed on 5 Jan 2024].
13. Osman, M. S., Abu-Mahfouz, A. M., & Page, P. R. (2018). A survey on data imputation techniques: Water distribution system as a use case. *IEEE Access*, *6*, 63279-63291.
14. Center for Behavioral Health Statistics and Quality. (2018). *Methods for Handling Missing Item Values in Regression Models Using the National Survey on Drug Use and Health (NSDUH): NSDUH Methodological Report*. Rockville, Maryland.
15. Howell, D. C. (2007). The treatment of missing data. *The Sage handbook of social science methodology*, 208-224.
16. Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC medical informatics and decision making*, *16*, 197-208.
17. Tutz, G., & Ramzan, S. (2015). Improved methods for the imputation of missing data by nearest neighbor methods. *Computational Statistics & Data Analysis*, *90*, 84-99..
18. What is the k-nearest neighbors' algorithm? IBM, Accessed: Jan. 06, 2024. [Online]. Available: https://www.ibm.com/topics/knn
19. Batista, G. E., and Monard, M. C. (2002). A study of k-nearest neighbor as an imputation method, *Front. Artif. Intell. Appl.*, vol. 87, pp. 251–260.
20. Luzi, O., & Grande, E. (2003). Regression trees in the context of inputation of item non-response: an experimental application on business data. Rivista di statistica ufficiale. Fascicolo 1, 2003, 1000-1030.
21. What is a Decision Tree | IBM. Accessed: Jan. 06, 2024. [Online]. Available: https://www.ibm.com/topics/decision-trees?mhsrc=ibmsearch_a&mhq=decision tree
22. What is Random Forest? | IBM. Accessed: Jan. 06, 2024. [Online]. Available: https://www.ibm.com/topics/random-forest?mhsrc=ibmsearch_a&mhq=random forest
23. Building a Random Forest Model: A Step-by-Step Guide. Accessed: Jan. 06, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#h-what-is-random-forest-algorithm
24. Song, I., Yang, Y., Im, J., Tong, T., Ceylan, H., & Cho, I. H. (2019). Impacts of fractional hot-deck imputation on learning and prediction of engineering data. *IEEE Transactions on Knowledge and Data Engineering*, *32*(12), 2363-2373.
25. Karmitsa, N., Taheri, S., Bagirov, A., & Mäkinen, P. (2020). Missing value imputation via clusterwise linear regression. *IEEE Transactions on Knowledge and Data Engineering*, *34*(4), 1889-1901.
26. Fang, F. (2016). Regression analysis with nonignorably missing covariates using surrogate data", Statistics and Its Interface, Vol. 9, p. 123–130.
27. Nagarajan, G., & Babu, L. D. (2022). Missing data imputation on biomedical data using deeply learned clustering and L2 regularized regression based on symmetric uncertainty. *Artificial Intelligence in Medicine*, *123*, 102214
28. Xu, D., Peng, H., Wei, C., Shang, X., & Li, H. (2021). Traffic state data imputation: An efficient generating method based on the graph aggregator. *IEEE Transactions on Intelligent Transportation Systems*, *23*(8), 13084-13093.
29. Fang, C., Zhao, X., Xu, Q., Feng, D., Wang, H., & Zhou, Y. (2020). Aggregator-based demand response mechanism for electric vehicles participating in peak regulation in valley time of receiving-end power grid. *Global Energy Interconnection*, *3*(5), 453-463.

30. Sierra-Porta, D. (2024). Assessing the impact of missing data on water quality index estimation: a machine learning approach. *Discover Water*, *4*(1), 11.
31. Song, J., Yang, Z., & Li, X. (2024). Missing data imputation model for dam health monitoring based on mode decomposition and deep learning. *Journal of Civil Structural Health Monitoring*, 1-14..
32. Ji, Z., Zhou, M., Wang, Q., & Huang, J. (2024). Predicting the International Roughness Index of JPCP and CRCP Rigid Pavement: A Random Forest (RF) Model Hybridized with Modified Beetle Antennae Search (MBAS) for Higher Accuracy. *CMES-Computer Modeling in Engineering & Sciences*, *139*(2).Zhigang, S., Wang, G., Li, P., Wang, H., Zhang, M., and Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Systems with Applications 237: 121549.
33. LEONARD¹, F. J., & DRAGOS, G. C. (2024). Examining the relationship between independent and dependent variables in the County Directorate for Sport and Youth management using econometric equations. *Journal of Physical Education & Sport*, *24*(2).
34. Saha, P., Marouf, Y., Pozzebon, H., Guergachi, A., Keshavjee, K., Noaeen, M., & Shakeri, Z. (2024). Predicting time to diabetes diagnosis using random survival forests. *MedRxiv*, 2024-02.
35. Çolak, A. B., Sindhu, T. N., Lone, S. A., Akhtar, M. T., & Shafiq, A. (2024). A comparative analysis of maximum likelihood estimation and artificial neural network modeling to assess electrical component reliability. *Quality and Reliability Engineering International*, *40*(1), 91-114.
36. Labuzzetta, C. J., Coulter, A. A., & Erickson, R. A. (2024). Comparing maximum likelihood and Bayesian methods for fitting hidden Markov models to multi-state capture-recapture data of invasive carp in the Illinois River. *Movement Ecology*, *12*(1), 2.
37. Awasthi, S., Singh, G., & Ahamad, N. (2024). Classifying electrical faults in a distribution system using k-nearest neighbor (knn) model in presence of multiple distributed generators. *Journal of The Institution of Engineers (India): Series B*, 1-14.
38. Amusa, L. B., & Hossana, T. (2024). An empirical comparison of some missing data treatments in PLS-SEM. *Plos one*, *19*(1), e0297037.
39. Mei, Z., Zhao, T., & Xie, X. (2024). Hierarchical fuzzy regression tree: A new gradient boosting approach to design a TSK fuzzy model. *Information Sciences*, *652*, 119740.
40. Cheng, M., Zhao, X., Dhimish, M., Qiu, W., & Niu, S. (2024). A Review of Data-driven Surrogate Models for Design Optimization of Electric Motors. *IEEE Transactions on Transportation Electrification*.
41. Bulat, M., Mirković, S., Gazivoda, N., Pejić, D., Urekar, M., & Antić, B. (2024). An improved algorithm for the estimation of the root mean square value as an optimal solution for commercial measurement equipment. *Microprocessors and Microsystems*, *106*, 10504
42. Xu, D., Peng, H., Tang, Y., & Guo, H. (2024). Hierarchical spatio-temporal graph convolutional neural networks for traffic data imputation. *Information Fusion*, *106*, 102292.
43. Fieberg, J. (2024). *Statistics for Ecologists: A Frequentist and Bayesian Treatment of Modern Regression Models*. University of Minnesota Libraries Publishing.
44. Giorgio, A. (2024). Project and Implementation of a Quantum Logic Gate Emulator on FPGA Using a Model-Based Design Approach. *IEEE Access*.
45. Hmood, M. Y., Al-Qazaz, Q. N., (2009). Comparing some methods for a single imputed of a missing observation in estimating nonparametric regression functionJournal of Economics and Administrative Sciences 15(53):223-235
46. Ali, L., (2017). Multi – Linear in Multiple Nonparametric Regression, Detection and Treatment Using Simulation, Journal of Economics and Administrative Sciences 23(101):495-503
47. Rashid, D. H., Hamza, S.K., (2016). Comparison some of methods wavelet estimation for non-parametric regression function with missing response variable at random, journal of Economics and Administrative Sciences, Volume 22, Issue 90, Pages 382-406
48. Kadhem, S. (2020). Comparison of weighted estimated method and proposed method (BEMW) for estimation of semi-parametric model under incomplete data. *Journal of Economics and Administrative Sciences*, *26*(120).
49. Rashid, D. H., and Hamza, S. K., (2016). Comparison of some indirect wave estimation methods for the prognostic narcissus function when there is a lack of biodiversity. Journal of Economics and Administrative Sciences, 22(90), 382-382