International Journal of Computer Science Engineering & Technology
Approved by
National Science Library (NSL)
National Institute of Science Communication and Information Resources (NISCAIR)
Council of Scientific and Industrial Research (CSIR)– New Delhi. India .
ISSN :2455-9091

NOVEL MULTIPLE IMPUTATION COMPARISON ON LINEAR REGRESSION, LOGISTIC REGRESSION, PREDICTIVE MEAN MATCHING ALGORITHM

[1] A. Nithya Rani M.C.A.,M.Phil., M.B.A., [2] Dr. Antony Selvdoss Davamani
[1]Assistant Professor,
[1] Dept of Computer Science, [2] Reader in Computer Science,
[1] C.M.S College of Science and Commerce, [2] NGM College, (Autonomous), Pollachi,
[1, 2] Coimbatore, Tamil Nadu, India,

ABSTRACT: Missing values present challenges in the analysis of data across many areas of research. Handling incomplete data incorrectly can lead to bias, over-confident intervals, and inaccurate inferences. One principled method of handling incomplete data is multiple imputations comparison on linear regression, logistic regression, predictive mean matching algorithm.The results show that neither the order, nor the number of imputations have significant impact on the bias, mean square error, or coverage, under this set of conditions. This work provides a baseline framework for more complex situations and more complex assumptions imposed on the missing values and classification of missing data.

Keywords: [Missing Data, Multiple Imputationh]

---

1. INTRODUCTION

Missing data are observations which exist however were not recorded or recorded and after that lost. In clinical examinations missing data regularly result from withdrawal, wearing down and misfortune to development. In different settings the missing data could be created through a coarsening plan. Fragmented data may emerge because of a few unique reasons including refusal, whittling down, estimation errors or just numbness about of the individual made inquiry. Regardless of what the reason is, missing observations is an issue that must be managed in every single measurable territory. The missing data mechanisms to be insignificant two conditions must be satisfied. In the first place, the missing observations must miss indiscriminately (MAR). Second, the parameters in the missing data process must be unmistakable from those in the data. The missing data design portrays which esteems in the data framework thatare really missing, and can help in the decision of strategy for taking care of the missing data. Missing data designs are typically separated into monotone (MMP) and discretionary missing examples (AMP). Figure 1 represented into Missing data patterns.

$$A=\begin{bmatrix} x & x & x & x & x \\ x & x & x & x & ? \\ x & x & x & ? & ? \\ x & x & ? & ? & ? \\ x & x & ? & ? & ? \end{bmatrix},\quad B=\begin{bmatrix} x & x & x & x & x \\ x & x & x & x & x \\ x & x & x & x & ? \\ x & x & x & x & ? \\ x & x & x & x & ? \end{bmatrix},\quad C=\begin{bmatrix} x & ? & x & x & x \\ x & x & ? & x & ? \\ ? & x & x & ? & x \\ x & x & ? & ? & x \\ x & x & ? & ? & ? \end{bmatrix}$$

Figure 1: Examples of missing data patterns. Rows correspond to units and columns to variables. Matrices A, B and C have MMP, UMP and AMP respectively

MMP may rearrange the analysis of the inadequate data as it might take into account the probability capacity to be factorized into factors for each square of cases with missing observations in similar factors, which would then be able to be amplified independently. Strategies built exclusively for MMP generally request less calculations than those planned likewise to deal with AMP. It might now and then even be worth considering expelling few observations or credit esteems for a few factors utilizing a subjective missing data strategy keeping in mind the end goal to make a data set with a "monotone" missing data design.

Missing Data Methods

Numerous throwing so as to miss data approaches disentangle the issue away data. In addition, discarding data can prompt estimates with bigger standard blunders because of lessened specimen size.

1.1 Complete-case analysis:

An immediate method to manage missing data is to bar them. In the regression setting, this generally means finish case analysis: excepting all units for which the outcome or any of the sources of info are missing. In R, this is done consequently for customary regressions (data centers with any missingness in the indicators or result are neglected by the regression). In Bugs, missing esteems in un exhibited data are not allowed, so these cases must be banished in R before sending the data to Bugs, or else the factors with missingness must be explicitly shown.

Two issues emerge with complete-case analysis:

1. In the event that the units with missing values vary systematically from the completely ob-served cases, with the complete-case analysis.
2. In the event that numerous variables are incorporated into a model, there might be not very many complete cases, so that the vast majority of the data would be disposed of for the purpose of a basic analysis.

1.2 Available-case analysis:

Open case analysis moreover develops when a master fundamentally forbids a variable or set of factors from the analysis in light of their missing-data rates now and again called "finish factors examinations". In a causal acceptance setting as with various forecast settings, this may provoke oversight of a variable that is vital to fulfill the suspicions fundamental for pined for causal translations.Imputation hypothesis is always making and thusly requires reliable regard for new data. There have been various speculations got a handle on by analysts to speak to missing data yet the lion's offer of them show a great deal of slant. Several the definitely comprehended endeavors to oversee missing data include: hot deck and cool deck imputation; list wise and combine clever erasure; mean imputation; regression imputation; last observation passed on forward; stochastic imputation; and different imputation.

2. LITERATURE SURVEY

Author	Year	Research Abstract Contribution
1. SandipSinharay, Hal S.Stern and Daniel Russel	2001	This article introduces the idea behind Multiple Imputation, discusses the advantages existing techniques for addressing missing data, describes how to do problems, reviews for software available to implement MI and discusses of a simulation study aimed at finding out how assumptions regarding the imputation model affect the parameter estimates provided by Multiple Imputations.
2. Fulufhelo Vincent Nelwamondo A	2009	The merits of both these techniques have been discussed at length in the literature, but have never been compared to each other. This thesis contributes to knowledge by firstly, conducting a comparative study of these two techniques. The significance of the difference in performance of the methods is presented. Secondly, predictive analysis methods suitable for the missing data problem are presented. The predictive analysis in this problem is aimed at determining if data in question are predictable and hence, to help in choosing the estimation techniques accordingly. Thirdly, a novel treatment of missing data for online condition monitoring problems is presented.
3. Benjamin M. Marlin	2008	This paper focuses on the problems of collaborative prediction with non-random missing data and classification with missing features. We begin by presenting and elaborating on the theory of missing data due to Little and Rubin. We place a particular emphasis on the missing at random assumption in the multivariate setting with arbitrary patterns of missing data. We derive inference and prediction methods in the presence of random missing data for a variety of probabilistic models including finite mixture models, Dirichlet process mixture models, and factor analysis.
4. Eng. Camelia Lemnaru (VidrighinBratu)	2011	The current thesis ascertains the problem statement and provides an analysis of existing approaches for the major theoretical problems tackled and, in some cases, also systematic empirical studies. Also, it proposes a series of novel methods for improving the behavior of traditional classifiers in such imperfect scenarios. In the data pre-processing step, the current thesis introduces an original global imputation method, based on non-missing data and a novel joint pre-processing methodology, which proposes an information exchange between data imputation and feature selection. Also, an original subset combination method for improving the stability of feature selection across different problems and

		providing an assessment of the baseline performance of feature selection in a new problem is presented.
5. Olanrewaju Michael Akande	2015	Evaluates the performance of several multiple imputation methods for categorical data, including multiple imputation by chained equations using generalized linear models, multiple imputation by chained equations using classification and regression trees and non-parametric Bayesian multiple imputation for categorical data. These data afford exploration of practical problems such as multicollinearity and large dimensions. This thesis highlights some advantages and limitations of each method compared to others. Finally, it provides suggestions on which method should be preferred, and conditions under which the suggestions hold.
6. Alexander Hapfelmeier	2012	Alternative ways to handle missing values are the application of imputation methods and complete case analysis. Yet it is unknown to what extent these approaches are able to provide sensible variable rankings and meaningful variable selections. Investigations showed that complete case analysis leads to inaccurate variable selection as it may inappropriately penalize the importance of fully observed variables. By contrast, the new importance measure decreases for variables with missing values and therefore causes selections that accurately react the information given in actual data situations. Multiple imputation leads to an assessment of a variable's importance and to selection frequencies that would be expected for data that was completely observed. In several performance evaluations the best prediction accuracy emerged from multiple imputations, closely followed by the application of surrogate splits.

3. PROPOSED WORK

3.1 NOVEL MULTIPLE IMPUTATION COMPARISON ON LINEAR REGRESSION, LOGISTIC REGRESSION, PREDICTIVE MEAN MATCHING ALGORITHM

This paper gives an integrated perspective of implementing Novel Imputation systems for multiple imputation procedures. Because of the importance of making the correct decision, better classification procedures are necessary for clinical decisions. The major objective of this paper is to implement and compare the proposed framework with three classification Simple Linear Regression model,Logistic regression, Predictive mean matching to build up an automated decision support framework for Multiple Imputation practice. The purpose was to decide an ideal classification mechanism for Multiple Imputation plans with high diagnostic accuracy. Distinctive classification algorithms were tried and benchmarked for their performance. The performance of the classification algorithms is illustrated on benchmark datasets.

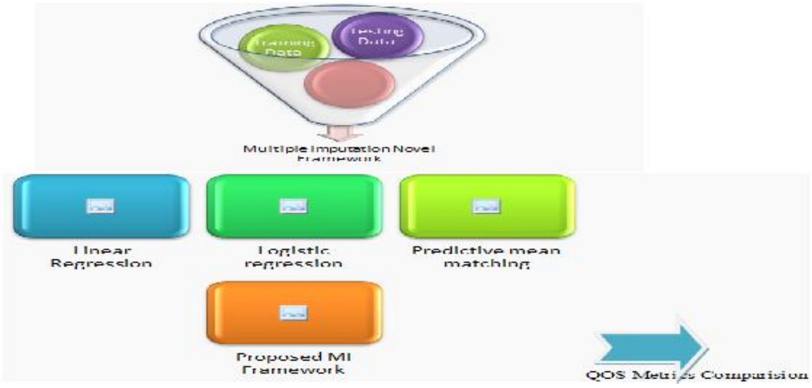

Figure 2: Proposed Overflow

A missingness structure is imposed as follows, 1.The first type of missing value is created with a missing totally at random structure to simulate a missing covariate. That is, a prespecified percentage of the values in X1 are randomly erased. Give MCAR% a chance to mean the percentage of missing values because of the first type of missingness. 2.The second type of missing value is created under a missing at random structure. Give MAR1% a chance to mean the percentage of missing values because of the first type of missing at random variable. Values in Y are evacuated in the event that they are above the best MAR1% percentile of X1. 3.The third type of missing value is created under a missing at random structure. Give MAR2% a chance to indicate the percentage of missing values because of the second type of missing at random variable. Values in Y are expelled on the off chance that they are beneath the base MAR2% percentile of X1.

The missing values are ascribed utilizing the standard package in R with varying numbers of imputations at each stage signified by the requested triple (L, M, N) and to such an extent that the request is MCAR%, MAR1%, MAR2%. Sider data has been utilized for the identification of missing data . For Training data 5-overlay cross approval model is utilized to test performances of the models. For a Sider dataset, all medications are randomly part into five subsets with parallel size. Each time, four subsets are consolidated as the preparation set, and whatever remains of the subset is used as the testing set.

4. EXPERIMENTAL RESULTS

The classification algorithm is a standout amongst the most vital capacities in the investigation of expansive datasets. Classification algorithms are the most generally utilized data mining models to separate profitable learning from gigantic measures of data (Dogan&Zuhal,2013). Classification is a data mining process that appoints things in an accumulation to target classifications or classes. The objective of classification is to foresee an objective class for each case in the dataset precisely. Numerous similar examinations are utilized to figure out which algorithm is most appropriate for a specific dataset. Classification ability relies upon the kinds of algorithms and the attributes of the data, for example, the level of imbalance, number of highlights, number of instances, and number of class composes. Besides, while missing values are dealt with by a specific imputation method, the classification algorithm is additionally influenced by the imputation method. In this manner, each extraordinary imputation method/classifier combine brings about an alternate execution, regardless of whether they treat similar data with the same missing values. Table 1 represented into comparison values of Novel Multiple Imputation Comparison on Linear Regression, Logistic Regression, Predictive Mean Matching Algorithm. Figure 3 represented into comparison of proposed overall metrics values.

	Linear Regression	Logistic Regression	Predictive Mean Matching Algorithm	Proposed Novel MI Framework
PCC	-0.6	-0.3	0.2	0.9
Mean Abs Sqr	0.3	0.1	0.5	0.9
RM Sqr Error	4	6	29	36
Precision	0.05	0.12	0.1	0.33
Recall	3	8	20	33
F-Score	4	15	19	38

Table 1: Comparison of proposed overall metrics values

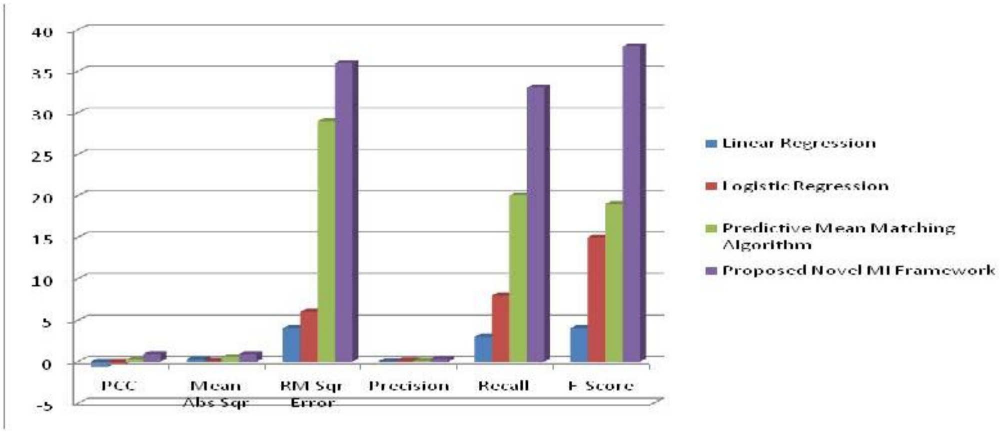

Figure 3: Comparison of proposed values

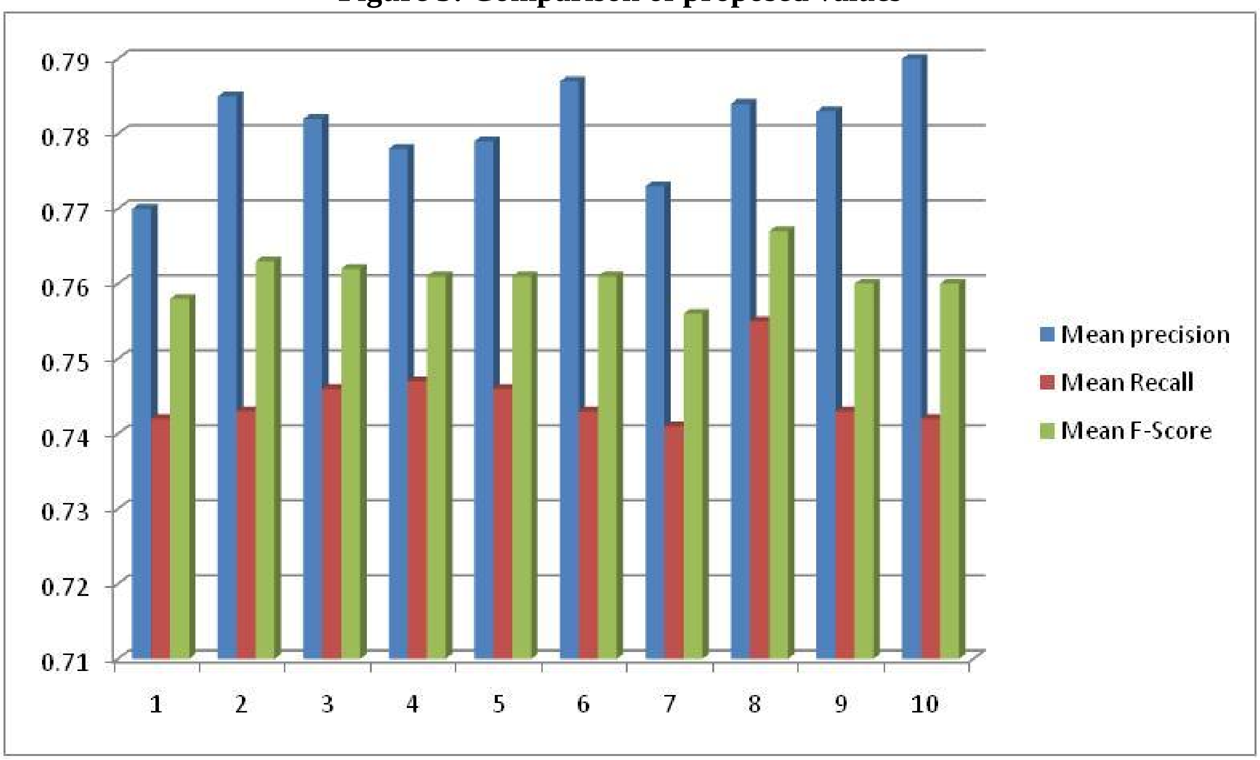

Figure 4: Comparison Mean metrics using 10 Runs values

CONCLUSION

This paper is experimented in an integrated view of implementing Novel Imputation systems for multiple imputation procedures. Because of the importance of making the right decision, better classification procedures are necessary for clinical decisions. The major outcome of this paper is to implement and compare the proposed framework with three classification

Simple Linear Regression model, Logistic regression, Predictive mean matching to develop an automated decision support system for Multiple Imputation practice. By experimenting the existing three classification models and the We determine an optimum classification mechanism for Multiple Imputation schemes with high diagnostic accuracy. Different classification algorithms were tested and benchmarked for their performance. The performance of the classification algorithms is illustrated on benchmark datasets.Proposed Novel MI Framework Mean Precision, Mean Recall, and Mean F-Score. The overall comparison of the above metrics results that the proposed novel MI Classification has shown significant improvement in identifying the missing values.

REFERENCES

[1] Andridge, R. R. (2011). Quantifying The Impact Of Fixed Effects Modelling Of Clusters In Multiple Imputation For Cluster Randomized Trials. Biometrical Journal, 53(1), 57- 74.

[2] Taljaard, M., Donner, A., &Klar, N. (2008). Imputation Strategies For Missing Continuous Outcomes In Cluster Randomized Trials. Biometrical Journal, 50(3), 329-345.

[3] VanBuuren, S. (2011). Multiple Imputation Of Multilevel Data. In J. K. Roberts & J. J. Hox (Eds.), The Handbook Of Advanced Multilevel Analysis (Pp. 173-196). New York: Routledge.

[4] Yucel, R. M. (2011). Random CovariancesAnd Mixed-Effects Models For Imputing Multivariate Multilevel Continuous Data. Statistical Modelling, 11(4), 351-370.

[5] Snijders, T. A. B., &Bosker, R. J. (2012). Multilevel Analysis (2nd Ed.). Thousand Oaks, California: Sage Publications.

[6] Keller, B. T., & Enders, C. (2014). A Latent Variable Chained Equations Approach For Multilevel Multiple Imputation. Paper Presented At The Modern Modeling Methods Conference, Storrs, Ct.

[7] Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., & Sterne, J. A. C. (2014). Joint Modelling Rationale For Chained Equations. Bmc Medical Research Methodology, 14(1), 28.

[8] Enders, C. K., Mistler, S. A., & Keller, B. T. (2014). Multilevel Multiple Imputation: A Review And Evaluation Of Joint Modeling And Chained Equations Imputation.

[9] De Gil, P. R., Pham, T., Rasmussen, P., Kellermann, A., Romano, J., Chen, Y.-H., &Kromrey, J. D. (2013). Gen_Eta2: A Sas® Macro For Computing The Generalized Eta-Squared Effect Size Associated With Analysis Of Variance Models. Paper Presented At The Sas Global Forum, San Francisco, Ca.

[10] Asparouhov, T., &Muthén, B. (2010f). Multiple Imputation WithMplus.