Exploring Missing Data using Adaptive LASSO Regression Imputation in Relation to Parkinson's disease

Qutaiba Humadi Mohammed, E.Srinivasa Reddy

Abstract: ***Parkinson's disease (PD) belongs to a class of chronic disorders that has degenerative neurological symptoms. In the clinical trials, different results falling in the areas of binary, ordinal, and continuous are analyzed to detect manifestation of the symptoms of this disease. A global test statistic is used to comprehensively evaluate the impact of all sorts of results. However, this disease predominantly faces the challenge of missing data that arise in the clinical results for varied reasons such as dropout, death, etc., therefore, imputation of such missing data must be carried out before conducting an intent-to-treat analysis. In fact, accuracy in data pertaining to disease progression may not be possible through statistical analysis without application of an appropriate mechanism that effectively handles missing data. In the p[resent paper, an Adaptive LASSO Imputation method has been proposed with its foundational basis on item response theory so that multiple imputations can be performed while dealing with multiple sources of correlation. The Root Mean Square Error (RMSE) formula was applied to evaluate the precision of each imputation method. The obtained results prove the better performance analysis of the proposed technique over all the known different algorithms.***

Index terms: HDD [High –Dimensional Data], Multiple Imputations, Regression, Missing Data

I. INTRODUCTION

Parkinson's disease (PD) falls under the category of chronic neurological disease with and degenerative manifestations. However, it is difficult to predict the primary factors causing the onset of this disease .But, research has shown a mixture of many environments related as well as genetic factors combine to cause PD. In common parlance, Parkinson's disease is widely known as a disorder affecting the central nervous system, which is caused when cells are lost from different domains of the brain. Such damaged cells contain a lot of nigra cells producing dopamine which is critical to ensure coordination of movement, besides transmitting messages chemically through signals throughout the brain.
When patients start losing such damaged cells, they are affected by movement disorders. When the literature data is reviewed, it becomes possible to perform clinical segregation of data from the various diagnoses done in PD. The selection is based on the traits denoting sensitivity and specificity of the representative clinical aspects. The clinical and pathological studies in patients deal with the investigation into the clinical, pathologic, and oncologic studies on the basis of incidence, characteristics as well as the prevalent risk factors. In PD, the symptoms are of two types- non-motor as well as motor. It is possible to make visual perception of the motor symptoms in patients, also known as cardinal symptoms. Such symptoms consist of resting tremor, tardy movement (bradykinesia), postural imbalance and stiffness. In fact, research has made it possible to observe non-motor symptoms in a specified time horizon, also known as dopamine-non-responsive symptoms. The other major symptoms like cognitive debility, sleep problems, sensory loss, constipation, speech and digestive issues, mysterious pain, drool, and low blood pressure. However, it is important to remember that none of such non-motor symptoms are final and definitive. But, accurate prediction of the onset of the disease is possible only aligned with other biomarkers from Cerebrospinal Fluid measurement (CSF) and dopamine transporter imaging. In the present work, the non-motor symptoms and the biomarkers like cerebrospinal fluid measurements and dopamine transporter imaging are taken as core factors in the analysis.

PD has manifestations like resting tremor, rigidity, slow movement, speech problems, swallowing problem, clustered as primary motor symptoms, while pain, depression are known as non-motor symptoms. The patients suffering from PD are examined by analyzing a total dataset and developing reliable and objective tools for assessment of PD. But, PD poses a challenge in the form of multi-dimensionality of data, besides the problem of missing data in vital parameters. Such issues are handled in a number of ways. This paper primarily seeks to deal with Parkinson's disease through its analysis of the missing data, preserving the crucial distinctiveness of the dataset, preserving the relationships between variables. In addition, lastly it estimates the missing values by applying the most appropriate method with the least error.

II. ORIGIN OF MISSING DATA

There are two types of Missing data viz., item non-response and unit non-response

Revised Manuscript Received on February 05 2019.
Qutaiba Humadi Mohammed, Research Scholar, Dept of CSE, ANUCOE&T, Acharya Nagarjuna University, A.P-India.
Dr.E.Srinivasa Reddy, Professor, Dept of CSE, ANUCOE&T, Acharya Nagarjuna University, A.P-India.

Unit Non Response

In case it is not possible to contact a respondent for completing a survey or one fails to return a questionnaire, then Unit non response is supposed to have happened [11][13]. It results in the respondent being no longer allowed to remain in the study. The unit non-response is managed by using weighting methods, such as Modeling Missing Data Imputation using Adaptive Lasso Imputation Method which is applied in case of Parkinson's data.

Item Non response

Mostly, it is possible to derive Item non response during the entire stretch self-administered questionnaires as well as after the applicant having finished his working survey. But, there may be certain missing options remaining to be filled or neglected while completing the portion of any document. When a participant fails to guess or answer any question, the missing values may also can surface, which also happens when an applicant forgets revert to a skipped question later. But, added to this, an applicant might get extra difficulty to respond to a long questionnaire if that particular case is nearer to the set due date. Also, in certain other cases, item non response involves missing data values, data lost in collection as well as during processing. The answer for a question can be raised in certain cases in which collection of data is not possible because of a failed device or a participant's simple forgetfulness. The extensive survey of handling item non-response have resulted in the need for application of weighting factor, imputation techniques and certain methods like EM algorithm. The study of this work completely handles only item non-response.

Missingness can be categorized in mathematical terms in three ways as elaborately defined by Little and Rubin as Missing completely at random (MCAR),Missing at random (MAR), and Missing not at random (MNAR) [11] [12] [13].

Missing Completely At Random (MCAR)

In case the missing value is not dependent on outcome and covariates, MCAR takes place, like the probability that an observation x_i is missing is not related to the value of x_i or to the value of any other variables. For instance, in survey research, the participant without any specific intention may flip over two pages in place of only one, thereby causing an entire page of missing observations. Little's MCAR test is regarded as most effective test that deals with missing data belonging to the type i.e. missing completely at random. In case the *p* value for Little's MCAR test is insignificant, MCAR represents the data, while missingness is excluded from the analysis[11][12]. The list wise deletion of values is deemed proper in case of a smaller number of missing values.

Missing At Random

In MAR, the missing observation does not depend on the measurement of the variable but is dependent upon another variable. For instance, the data meets the requirement of the missingness as independent of the value of x_i after controlling another variable. Missing at random takes place in case MCAR is not indicated by Little's MCAR test as significant, while there is a chance to predict missingness by other observed variables, independent of any other unnoticed variables [14][15]. In MAR, it is possible to predict missingness from noticed variables and justify the application of multiple imputations (MI).

Missing not at random (MNAR)

Missing values are classified as missing not at random (MNAR) if they do belong to neither MCAR nor MAR. MNAR, referred other wisely as non-ignorable assume significance in case missing values depend on non-observed data even after controlling all the noticed data. On the basis of external research design values are imputed in case the variable with missing data has no sufficient correlation with rest of the variables in the dataset, which is a unique way to deal with MNAR (non-ignorable missingness)[10][17][4]

III. RESEARCH BACKGROUND

Multiple Imputations

Multiple imputations belong to a class of statistically significant methodologies that offer better solutions to missing data problems. Such a simple technique produces the right balance to the original data. The missing values for any variable are predicted in multiple imputations by applying the known values from the other variables. The predicted values represented as imputes are used in place of the missing values that result in developing a full data set, known as an imputed data set. Such a method is performed many times and produces several imputed data sets. The standard statistical analysis is carried out on each imputed data set, producing multiple analysis results. Each imputed data set is assigned multiple imputation accounts for missing data through restoration of not just the natural unevenness within the missing data, but adding more uncertainties resulting from the estimation of missing data[12][14]. The maintenance of the original variability of the missing data gets complemented through creation of imputed values based on variables having relation to the missing data as well as ones that cause missingness. Entirely varied versions of the missing data are applied to make up for uncertainty, besides observing the variability between imputed data sets [15] [16]. In fact, it may be significantly noticed that imputed values endowed by imputation model do not intentionally predict the behavior of a specific missing value. Rather, such a modeling method is expected to form an imputed data set for maintaining the overall variability within the population, besides preserving the correlation with other variables. In this process, the multiple imputations provides a researcher the ability to conserve crucial feature of the dataset in all their entirety, for instance, means, variances, regression parameters.

The degree of suitability of such techniques is dependent upon the relation of missing data to other variables, which necessitates the need to enjoin multiple imputation method with additive least absolute shrinkage and selection operator (LASSO) for imputing data into the dataset. The simulated results of House Price data set are presented after examining with several missing data. The present study has

conclusively proved that the conjoined application of multiple imputation and additive LASSO model created the most promising experimental results.

Additive LASSO model

Assume that in linear regression with Q the $n\times 1$ response vectors with q_i and X the $n\times xp$ predictor matrix with row vector x_i with i^{th} subject values for the i^{th} subject. Lasso Regression model assumes that either of the observations which are independent or they can have q_i which are provisionally independent given with the values x_{ij} where they x_{ij} have been consistent so that $\sum_i x_{ij}/n=0$ and $\sum_i x_{ij}^2/n=1$ In addition, the value q_i have been inclined to center to have sample mean 0 under these considerations, the lasso estimates are given by

$$\beta=\arg\min_{\beta}\sum_{i=1}^{n}(q_i-\beta_j^{'}x_i)^2 \quad \text{subject to}$$

$\sum_{j=1}^{p}|\beta_i|\le t$ where $t\ge 0$ is a tuning value which will controls the amount of shrinkage applied to the estimated parameters and, therefore, the degree of variable selection

Coefficient estimates can be calculated

Let be the A "active set" of covariates most correlated with the "current" residual

Initially $A=\{x_{j1}\}$ for some covariate x_{j1}

Take the largest possible step in the direction of x_{j1} until another covariate x_{j2} enters A

Continue in the direction equiangular between x_{j1} and x_{j2} until third covariate x_{j3} enters A

Continue in the direction equiangular between x_{j1}, x_{j2}, x_{j3} until a fourth covariate x_{j4} enters A

This procedure continues until all covariates are added at which to the missing point.

$$\beta=\begin{cases} c_j+\lambda/a_j & c_j\leftarrow\lambda \\ 0 & c_j\in[-\lambda,\lambda] \\ (c_j-\lambda)/a_j & c_j>\lambda \end{cases}$$

IV. RELATED WORK

This study has been implemented on Parkinson's Data Set which is retrieved from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/parkinsons). This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals

Exploring Missing Data

3.1.1 Visualization methods for missing values

The Fig.1 shows the missingness in Parkinson Data. The nms_d3 is having high missing percentage of 25% where as rigidity, nms_d7, nms_d4 having 20% of missingness

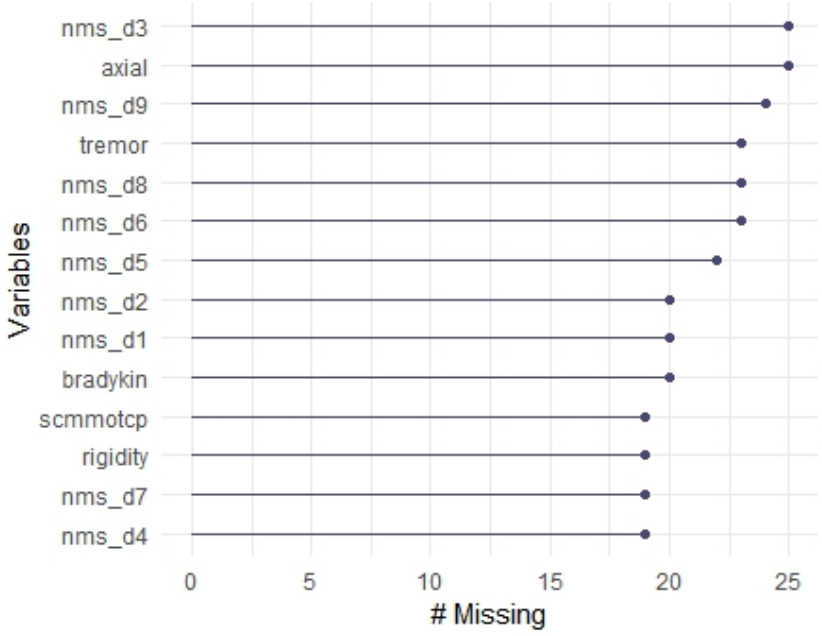


Fig. 1 Percentage of Missingness in Parkinson Data

Aggregation Plot

The plot is used to locate the distribution of missing values at MAR that provides an overview of the missing

data with missing values patterns. It also examines in detail the location-related aspects of missing samples as well as their frequency. In Figure 3a, the outcomes of missing data exploration on absolute frequency and proportion of each variable in dataset is included. However, the description about observations with missingness is provided in the plot in left side. At right side, the projection is done by combining red and yellow boxes, representing the missingness in variables. But, the frequency of missingness in each variable combination is derived in the plot as projected in a separate box. The plot is found to be enhancing missingness pattern when aligned with linear order of frequency. In similar way, the left bottom row carries the observations with missing values much the same as top left. The distribution of missingness every variables includes twenty fifth in nms_d3 and axial and therefore the reaming information is shown within the Fig1. However, the study during this work limits missingness distribution at the most. With relevancy the proportional results discovered by the plot highest frequent combination shown in variable nms_d3 to lowest frequent combination in variable nms_d4.

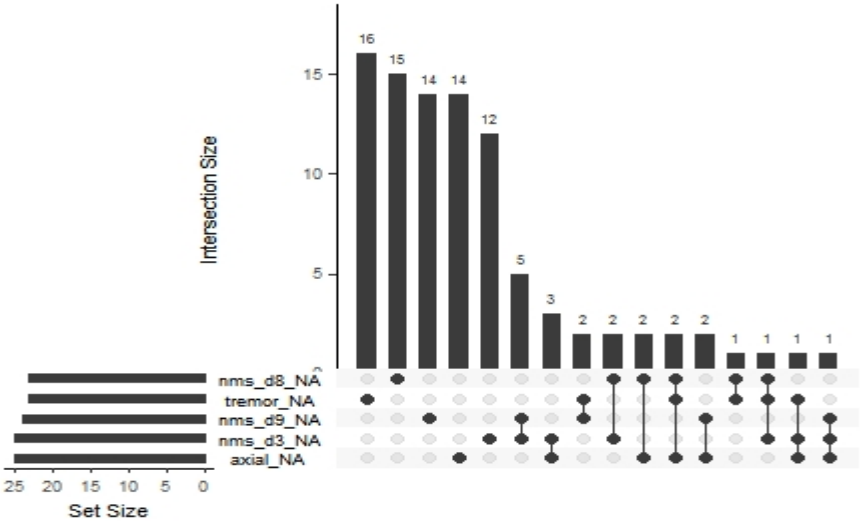


Fig. 2 Percentage of Missingness in Parkinson Data

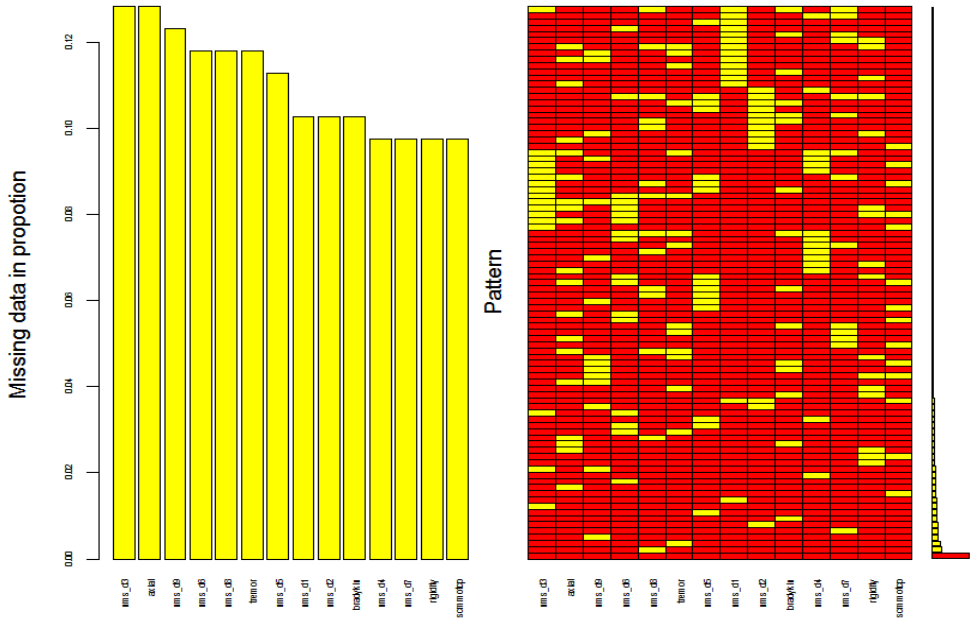


Fig. 3 Aggregation Plot for Exploring Missingness in Parkinson Data

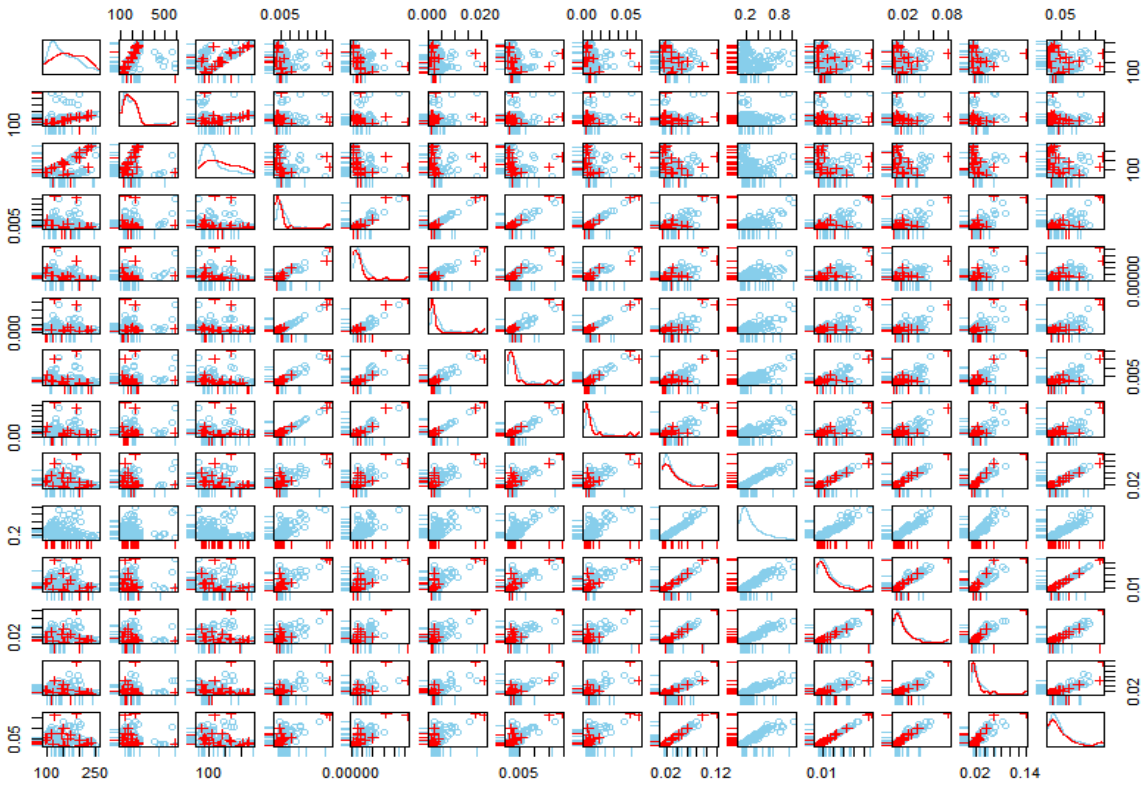
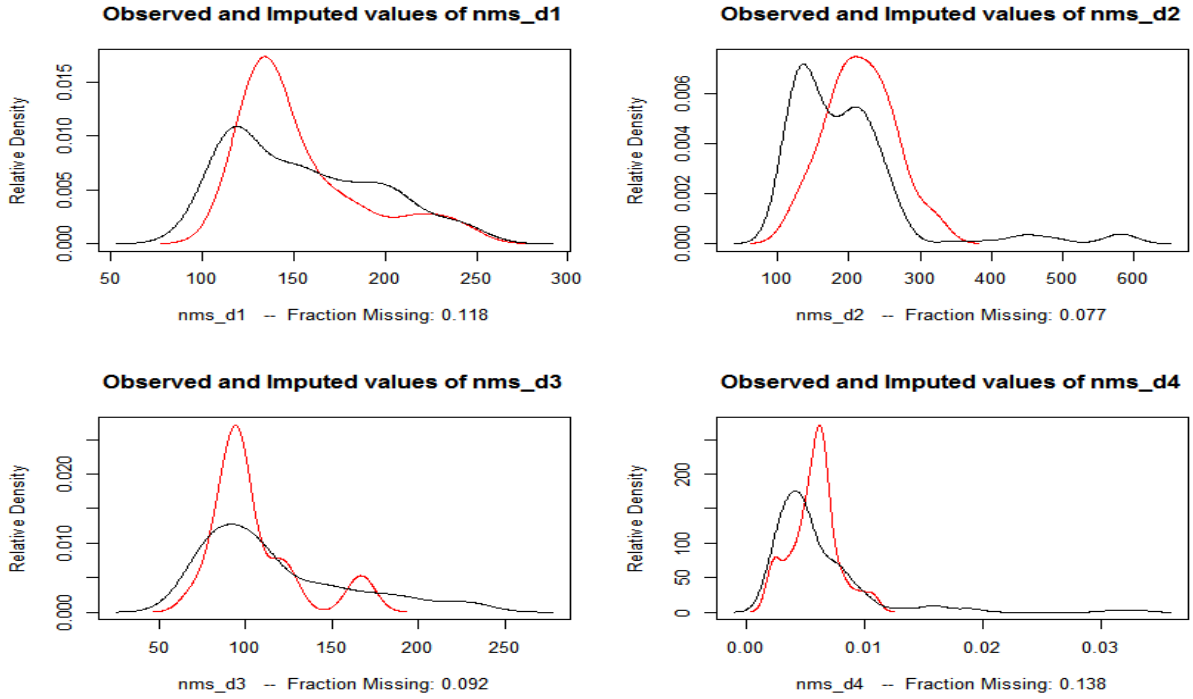
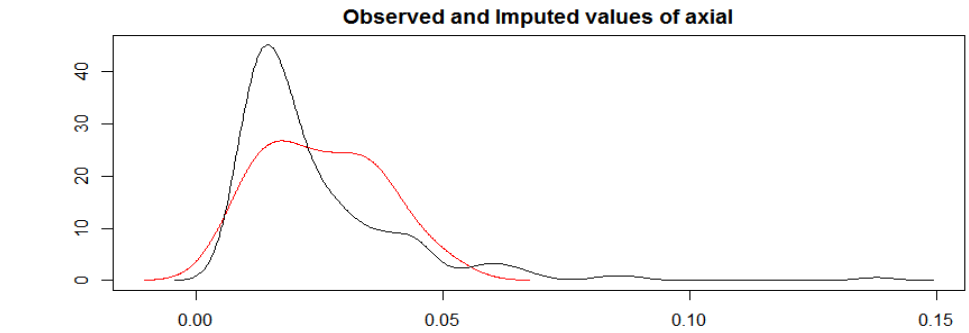


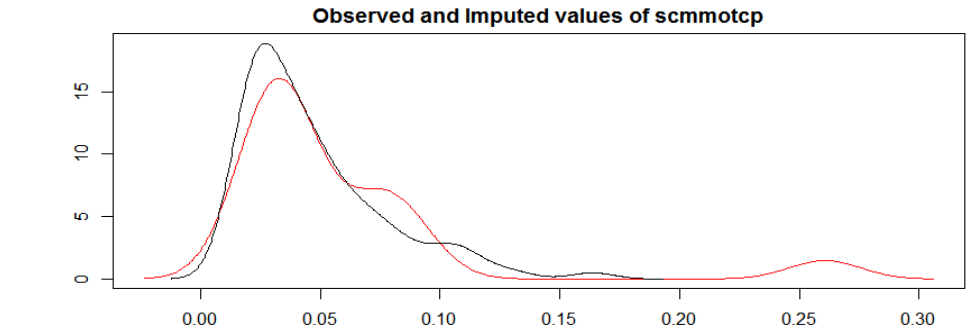
Fig. 4 Scatter Plot of Parkinson Data showing missing data highlighting tremor symptom

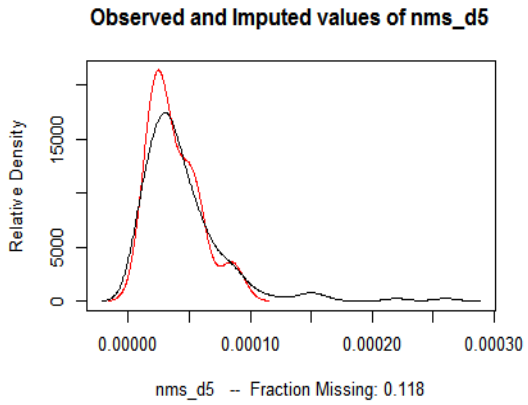
Adaptive LASSO Imputation

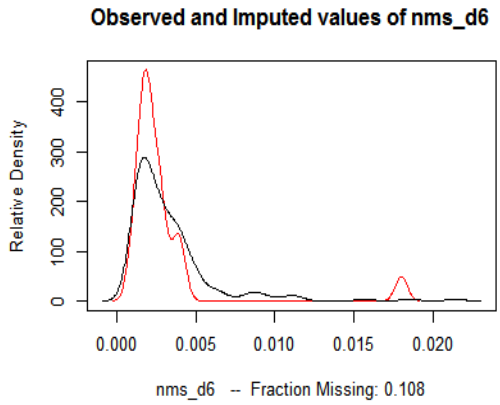
The Fig 5, Fig 6 and Fig 7 shows imputation efficiency by using Adaptive LASSO Imputation when compared with observed variable vs imputed value. The imputation variables include nms_d1,nms_d2, nms_d3, nms_d4, nms_d5, etc. It is clearly visible that Adaptive LASSO imputation can recommended to for imputation in large data

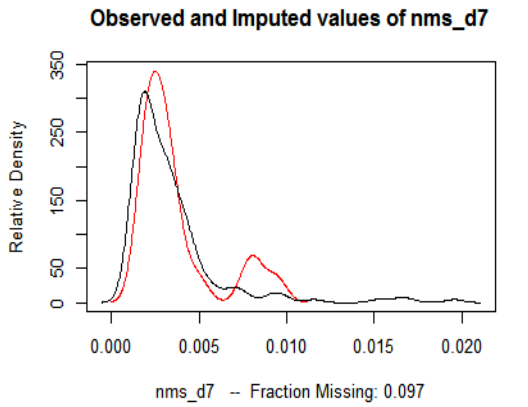


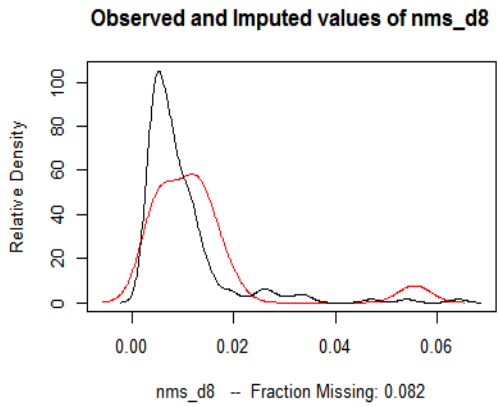

Observed and Imputed values of axial
0
10
20
30
40
0.00
0.05
0.10
0.15


Observed and Imputed values of scmmotcp
0
5
10
15
0.00
0.05
0.10
0.15
0.20
0.25
0.30


Observed and Imputed values of nms_d5
Relative Density
0
5000
15000
0.00000
0.00010
0.00020
0.00030
nms_d5 -- Fraction Missing: 0.118


Observed and Imputed values of nms_d6
Relative Density
0
100
200
300
400
0.000
0.005
0.010
0.015
0.020
nms_d6 -- Fraction Missing: 0.108


Observed and Imputed values of nms_d7
Relative Density
0
50
150
250
350
0.000
0.005
0.010
0.015
0.020
nms_d7 -- Fraction Missing: 0.097


Observed and Imputed values of nms_d8
Relative Density
0
20
40
60
80
100
0.00
0.02
0.04
0.06
nms_d8 -- Fraction Missing: 0.082

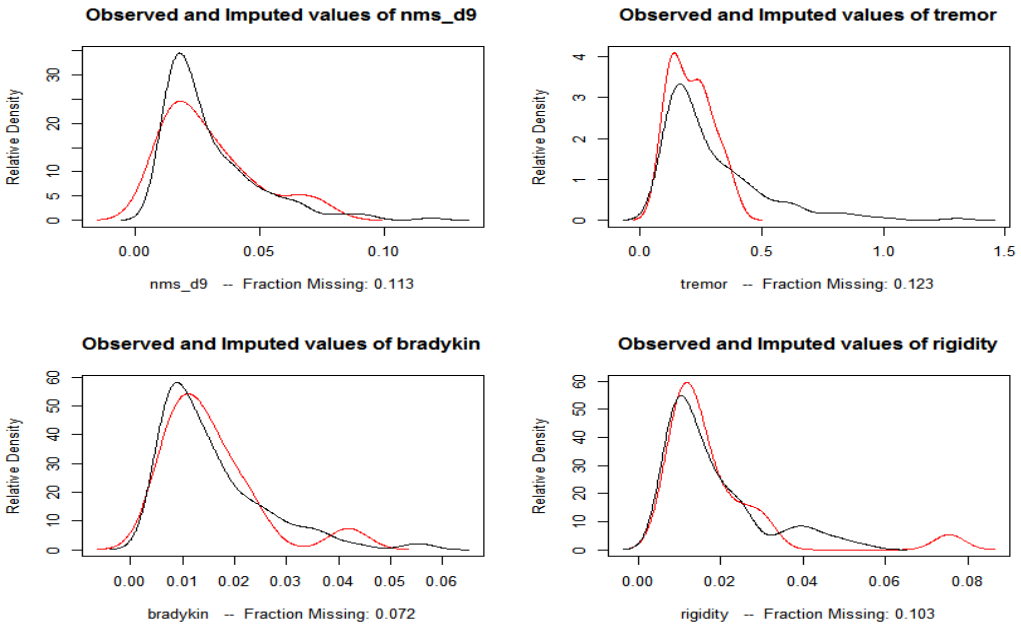


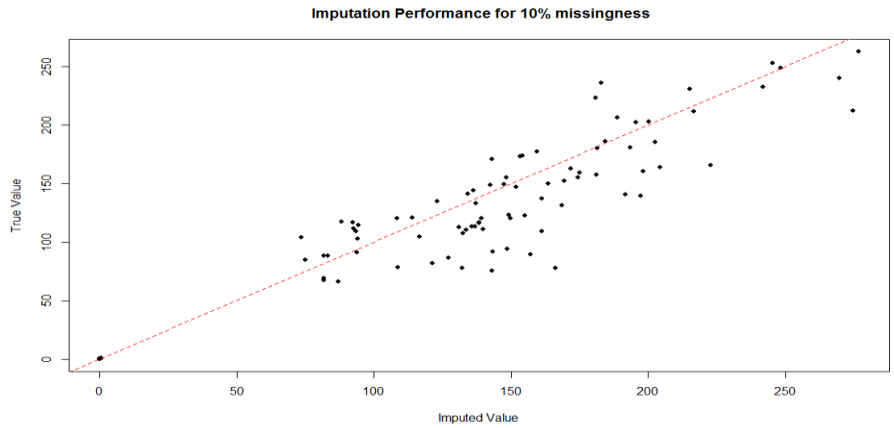
Table.1 Adaptive LASSO Imputation analysis of dataset

Observation	Estimate	Standard Error Rate	t_value	Pr>|t|	F_value	Pr(>F)
nms_d1	-0.000745	0.00164	12.792	1.4e-35	328.996	< 2e-16
nms_d2	-2928.253	1036.944	-2.824	4.8e-03	233.270	< 2e-16
nms_d3	1042.229	80.976	12.871	5.6e-36	387.957	< 2e-16
nms_d4	27877.938	2508.624	-11.113	1.4e-27	125.372	< 2e-16
nms_d5	3789.322	3558.773	1.065	2.9e-01	1.134	0.2872
nms_d6	-8062.642	3004.422	-2.684	7.4e-03	6.084	0.0138

This imputation is also compared with various other imputation methods with varying 10%, 30% and %50 missingness

Table. 2 RMSE value for different types Imputations

Observation	Adaptive LASSO	Ridge	StepBothR	StepForR	StepBackR
10%	0.2327802	0.2360795	0.2388631	0.2393628	0.2406279
30%	0.3212882	0.3432687	0.3654264	0.3898151	0.4105957
50%	0.4256324	0.4479363	0.4698911	0.4831899	0.5232519



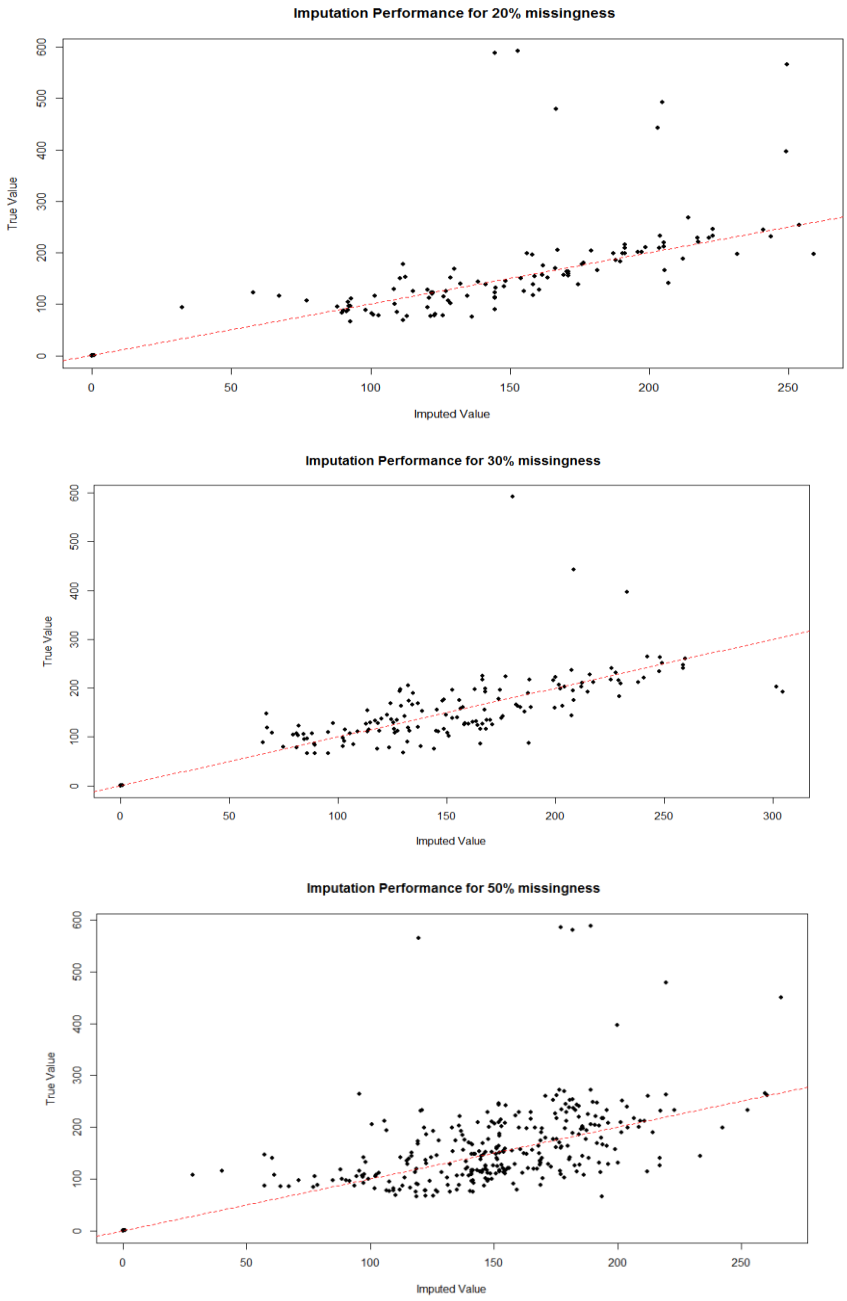


Fig.6 Imputation results for different percentage of missingness

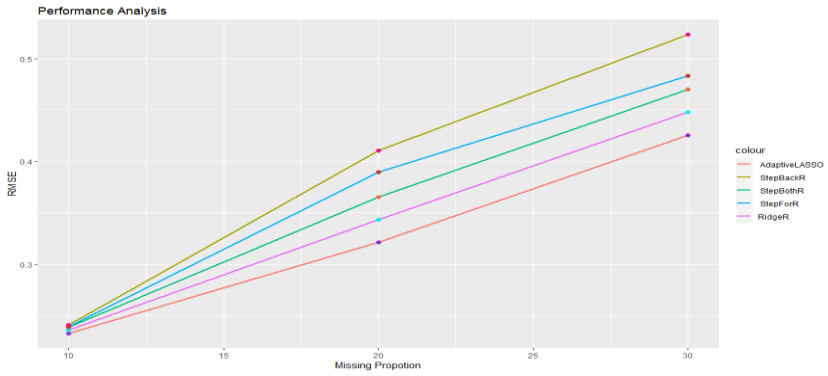


Fig.7 Comparison of RMSE with different Imputation methods

V. CONCLUSION

In the article we are apply 3- imputation techniques Mean, predictive mean and additive LASSO are employed. Finally results show imputations by additive LASSO is preferred multiple imputation technique.

REFERENCES

1. Barnard, J. and Meng, X.L. (1999). Applications of multiple imputation in medical studies: from AIDS to NHANES, Statistical Methods in Medical Research, 8, 17–36.
2. Bennett, D.A. (2001). How can I deal with missing data in my study? Australian and New Zealand Journal of Public Health, 25, 464–469.
3. Durrant, G.B. (2005). Imputation methods for handling item-nonresponse in the social sciences: a methodological review, ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute. NCRM Methods Review Papers NCRM/002.
4. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. Journal of clinical epidemiology. 2006;59(10):1087-91. doi: 10.1016/j.jclinepi.2006.01.014.
5. Horton, N.J. and Lipsitz, S.R. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing variables, American Statistical Association, 55, 244–254.
6. Janssen, K.J.M., Donders, A.R.T., Harrell Jr., F.E., Vergouwe, Y., Chen, Q., Grobbee, D.E. and Moons, K.G.M. (2010). Missing covariate data in medical research: to impute is better than to ignore, Journal of Clinical Epidemiology, 63, 721–727.
7. 7 Little, R.J.A. and Rubin, D.B. (2002). Statistical analysis with missing data, 2nd ed., New York: John Wiley and Sons, Inc., 381 pages.
8. Little, R.J.A. (1988). A test of missing completely at random for multivariate data with missing values, Journal of American Statistical Association, 83, 1198–1202.
9. Little, R. (2011). Calibrated Bayes, for Statistics in general, and missing data in particular, Statistical Science, 26, 162–174.
10. Popov, S. (2006). Large-scale data visualization with missing values, Technological and Economic Development of Economy, 12, 44-49..
11. Rubin, D.B. (1987). Multiple imputation for nonresponse in surveys, 1st ed., New York: John Wiley and Sons, Inc., 258 pages.
12. Rubin, D.B. and Schenker, N. (1991). Multiple imputation in health-care databases: An overview and some applications, Statistics in Medicine, 10, 585–598.