A Multi Linear Regression Approach for Handling Missing Values with Unknown Dependent Variable (MLRMUD)

A. K. Alhebshi
Department of Computer Engineering
Faculty of engineering
Cairo University
Cairo, Egypt
engcmp505@gmail.com

M.F. Ahmed
Assistant professor in Computer Engineering
Faculty of engineering
Cairo University
Cairo, Egypt
mona_farouk@eng.cu.edu.eg

A. F. Atiya
Professor in Computer Engineering
Faculty of engineering
Cairo University
Cairo, Egypt
amir@alumni.caltech.edu

***Abstract*—many problems in data applications are plagued with missing values. The Missing Value problem (MV) is the problem of predicting these missing values, in an attempt to make full use of the data. Simply deleting the missing record will waste precious information. In this work a new approach is proposed, the so-called MLRMUD. It is based on Multiple Linear Regression is used to predict Missing values for a data set with Unknown Dependent variable. It is applicable if complete rows are at least 20%. If they are less than that the Mean method is used to fill some rows until the complete rows reach 20%. After that MLRMUD can be applied normally. This approach is composed of three algorithms; splitting algorithm, dependent variable selection algorithm and multi-linear regression algorithm. MLRMUD is compared to other methods in the literature where it is proved that it outperforms them all in the accuracy of missing value computation determined in terms of to Root Mean Square Error (RMSE) and Mean Standard Error (MSE). A method to determine the unknown, dependent variable from the training set is proposed. The results show that the proposed method can successfully select the dependent variable with an accuracy of 83% over all the datasets examined.**

Keywords— missing values, splitting algorithm, dependent variable, multi linear regression, regression coefficients

I. INTRODUCTION

Many existing data sets contain missing values. These could impact the data analysis in a negative way. The data sets contain missing values due to various reasons: manual data entry procedures, lapses in data collection, incorrect measurement and equipment errors. The missing values can make it very difficult to use data analysis methods which require complete data. These missing values should be predicted to be able to use data analysis methods accurately. The work done always assumes the knowledge of which variables are the dependent variables and which are the independent variables. There are some real applications where the identity of the dependent variable is unknown by nature. These applications include network traffic prediction [1]. In this application each node is assumed to have a specific traffic. Moreover, each nodes' traffic may influence and may be influenced by the other node's traffic. So it is not known which variables are the independent variables (influencers), and the dependent variables (being influenced). Another important application is the weather prediction problem. In this application the temperature is determined by a certain number of weather stations and it is required to know which are the dependent stations and which are independent. (The independent stations will be influenced, and are typically at the forefront of heat/cold waves.). This work handles this issue, and develops an approach that can estimate which are the dependent versus independent variables. Moreover, the proposed work also provides an effective missing values prediction approach. The proposed approach produces a higher prediction accuracy, as compared to some existing algorithms.

A. Problem Statement

Let G (MV, UDV) be a data set where MV is Missing Value, UDV is Unknown Dependent Variable. The Multi-linear regression equation is given as:

$$Y = b0+b1X1 + b2X2 +\ldots + box \quad (1)$$

Where **Y** is the dependent variable and $X_{i\ (I:\ 1\ to\ k)}$ are the independent variables and $b_{i\ (i:\ 0\ to\ k)}$ are the regression coefficients. The missing values for this problem should be of numeric type. The dependent variable value depends on the values of the independent variables; the level of relation between the dependent variable and the independent variables varies. In figure 1, the input data set G (MV, UDV) has missing values and unknown dependent variable. The dependent variable selection algorithm is applied to the data set G (MV, UDV) to infer the dependent variable. G (MV) is

the dataset with missing values after solving one of its unknown parameters which is the dependent variable. Multi linear regression Algorithm is applied to the data set G (MV) to predict missing values resulting in the complete data set G

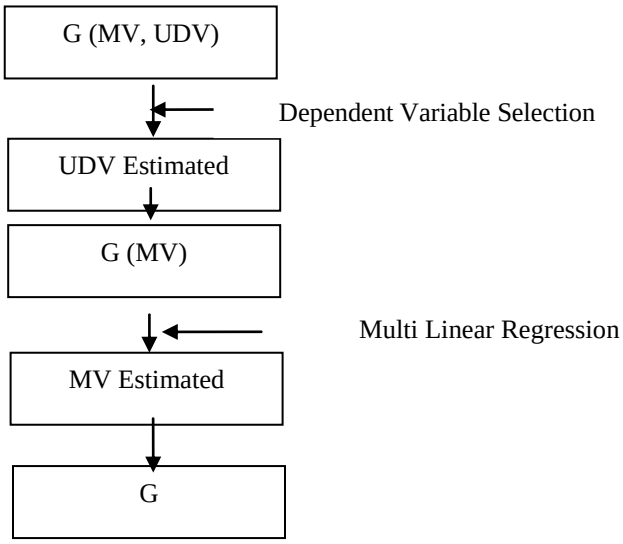

Fig. 1. Missing value problem with unknown, dependent variable

II. RELATED WORK

This section summarizes the work done to handle missing values in data sets.-Handling missing values depends on the type of relation that best describes the data set attributes which can be linear or nonlinear.

A. Nonlinear Data Set Algorrithms

Generally the data can be grouped into two main types: numerical and categorical. Numerical missing values are easier to predict while categorical value problems are complex and contain multiple types. The reason is that the relations among the categorical values of the different attributes are hard to model.

There is much work done on the categorical missing value problem**.** Kaiser et al [2] proposed a new technique which is based on association rules. It splits the data set into a training set and a test set. Association rules are applied in the training set to learn the categorical values and then predict the missing values in the test set. Ferrari et al [3] proposed a forward imputation procedure which is used in the context of the Nonlinear Principal Component Analysis. This algorithm is used to obtain indicators from a large dataset.

On the other hand, even more work was also done to deal with numerical missing values. Pan et al [4] developed two algorithms: KNN impute and ARL impute. Both algorithms are used to estimate missing values in data sets based on imputation. KNN gives the best result only when the k value is between 10 and 20. This KNN impute algorithm is used for any particular column, but the ARL imputation algorithm is used for finding many missing values from the same column. AP et al [5] proposed Expectation-Maximization approach which is of maximum likelihood that can be used to create a new data set. This approach begins with the expectation step where some parameters are estimated (i.e. Variance, mean). These parameters are then used to create models of linear regression which are used to fill missing values. The expectation step is then repeated with other values of parameters. The operation continues to get better predicted missing values. P.saravanan et al [6] proposed a new technique to deal with missing values. The approach is called Fuzzy Possibilistic C Means Algorithm (FPCM). It is a combination of possibilistic C Means algorithm and fuzzy C Means algorithm. It has the advantages of possibilistic C Means such as handling noisy data and that of fuzzy C Means algorithm such as the possibility that the data can belong to more than one cluster. This gives the best result for overlapped data. The FPCM outperforms the FCM approach. Chhabra et al [7] compared six different approaches with respect to the accuracy of missing value computation; these approaches are Predictive Mean Matching, Multiple Random Forest Regression Imputation, Multiple Bayesian Regression Imputation, and Multiple Linear Regression using non-Bayesian Imputation, Multiple Classifications and Regression Tree (CART), Multiple Linear Regression with Bootstrap Imputation. Multiple Bayesian Regression Imputation is the best in accurately. The work by [8] used the concept of associative memories for missing values. Hassan et al [9] used the concept of ensemble neural networks for handling missing values. Mohamed et al [10] considered missing values for time series forecasting applications, where the forecasting is performed using exponential smoothing [11].

B. Linear Data Set Algorithms

Regression imputation assumes that the value of one variable changes in some linear way with other variables. The missing data are replaced using a linear regression function. This method depends on the assumption of a linear relationship between attributes. Raval et al [12] proposed a new technique to handle missing values for numeric attributes using simple linear regression. It divides data into training and testing. Test data contain 20 % missing values. The algorithm does not deal with training set at all. It predicts the missing values in the test set directly not dependent on learning on the training set, so no need to access all the data. This method uses linear equation 2.

$$Y = X + e \qquad (2)$$

Assuming the regression coefficients b0 = 0, b1 = 1. Where Y is the dependent variable that has missing values, X is the independent variable that is the summation of the previous values of dependent variable Y and e is the error which equals

the sum of previous values/ number of previous values. Shelke et al [13] proposed new technique which uses a conceptual data reconstruction by using statistical models of multiple linear regressions to predict missing values in data sets. Hailin et al [14] also proposed a new technique to handle missing values in linear data set using equation 3

$$X^*_i = a_0 + \sum_{j=1}^{p} a_j X_{i-j} \quad (3)$$

It calculates the value of p by using covariant coefficient, which reflects the relationship between random variables. If the covariant coefficient is too small it ignores the relationship. Beyad et al [15] proposed a new way to handle missing values in linear data by using linear regression in terms of matrices as shown in equation 4

$$D = C*A+R \quad (4)$$

Matrix D is treated as the dependent variables, matrix C as regression coefficients, the matrix A as the independent variables and matrix R as the residual error.

III. Proposed Algorithm

This work proposes a new approach for handling missing values assuming a linear relationship between the data set attributes. The proposed approach, MLRMUD, has four main steps: Splitting the data set into training set and test set; deriving the unknown, dependent variable from the training set; using the training set with the derived, dependent variable to create a linear regression model; and finally, using the learnt model to predict missing values in the test set. If the data set contains < 20 % complete rows, the Mean is used as a way to fill data until the threshold of MLRMUD is satisfied (≥20% complete rows). This step improves the accuracy of the procedure. Figure 2 shows MLRMUD flowchart

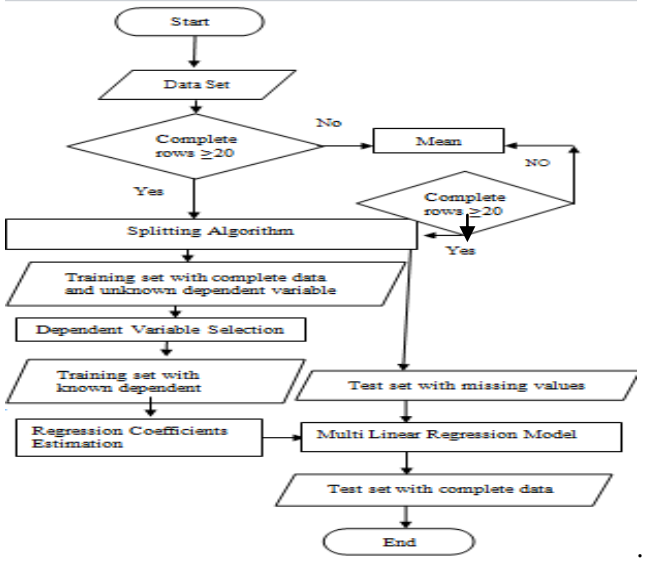

Fig. 2. MLRMUD flowchart

A. The Proposed Splitting Algorithm

Let the data set be G. If G contains complete data rows at least 20%, then split G in Training set and Test sets such that the Training set contains all full rows and the Test set contains rows with MV. Otherwise, the Mean method can be used to fill a number of rows until the percentage of complete rows equals 20%.

TABLE I. An example showing the proposed splitting algorithm

Data Set

X1	X2	X3	X4	X5
10	50	12		100
25	18	57	98	30
15	79	78	20	
100	90	80	70	60
45	54	68	99	75
12		58	65	

Training Set

X1	X2	X3	X4	X5
25	18	57	98	30
100	90	80	70	60
45	54	68	99	75

Test set

X1	X2	X3	X4	X5
10	50	12		100
15	79	78	20	
12		58	65	

B. Dependent Variable Selection Algorithm

Assume the training set has n attributes X_1, X_2... Xn
The selection algorithm is described in figure 3 below

Loop from i=1 to n -1
Loop from j = i+1 to n
CorrelationV = Correlation_test_P-value (X_i, X_j)
If CorrelationV <= 0.05
Add Xj to List in the location corresponding to Xi
End loop
End loop

Fig. 3. The dependent variable selection algorithm

The dependent variable is the attribute that is the most repeated in the List as it has many other attributes which are correlated to it.
The algorithm can be explained as follows. If the p-value of the correlation is less than 0.05, then there is a statistically significant correlation among the variables. The variable that has the most correlations with other variables is likely the dependent variable, because it has more relations that depend on other variables.

C. Regression Model Formation

The least squares multipliers method is used to find the coefficients for the linear regression model represented in equation 5 below.

$$Y = B_0+B_1X_1+B_2X_2+\ldots+B_K X_K + E \quad (5)$$

Let each of the k independent variables, x_1, x_2... x_k, has n levels. Then x_{ij} represents the i th level of the j th independent variable x_j and y1, y_2... y_n, have n levels. This can be expressed in the following way:

$$y_1 = b_0+b_1x_{11}+b_2x_{12}+\ldots+b_k x_{1k} + e_1 \quad (6)$$

$$y_2 =b_0+b_1x_{21}+b_2x_{22}+\ldots+b_k x_{2k} + e_2 \quad (7)$$

$$y_i = b_0+b_1x_{i1}+b_2 x_{i2}+\ldots+b_k x_{ik} + e_I \quad (8)$$

$$\ldots\ldots\ldots$$

$$y_n = b_0+b_1x_{n1}+b_2 x_{n2}+\ldots+b_k x_{nk} + e_n \quad (9)$$

Equation (5) is reformatted to

$$Y = X B + E \quad (10)$$

By using matrix notation

$$Y= \begin{vmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{vmatrix} \quad X= \begin{vmatrix} 1 & x_{11} & x_{12} \ldots & x_{1n} \\ 1 & x_{21} & x_{22} \ldots & x_{2n} \\ \ldots\ldots\ldots\ldots\ldots & & & \\ \ldots\ldots\ldots\ldots\ldots & & & \\ \ldots\ldots\ldots\ldots\ldots & & & \\ 1 & x_{n1} & x_{n2}\ldots x_{nn} & \end{vmatrix} \quad (11)$$

The first column in matrix X is all ones corresponding to **B**0.
From equation 12 after applying
X*B will give the same results of equations 8, 9, 10, 11

$$B = \begin{vmatrix} B_0 \\ B_1 \\ . \\ . \\ . \\ B_n \end{vmatrix} \quad E = \begin{vmatrix} E_0 \\ E_1 \\ . \\ . \\ . \\ E_n \end{vmatrix} \quad (12)$$

The regression coefficient formula is

$$B = (X^{/}X)^{-1} X^{/} y \quad (13)$$

B: Regression coefficient matrix
X: Independent variable matrix
y: Dependent variable matrix
$X^{/:}$ Transport matrix of X

D. The Complete Proposed Algorithm

If the data set has < 20% complete rows, the Mean method is initially used to fill the data until the threshold of MLRMUD is satisfied (≥20% complete rows) otherwise MLRMUD is directly applied as follows:

1. Split the data set into the training set and the test set using the splitting algorithm. The test data contains at least one missing value in each row. The training data contain complete rows.
2. Use the training data to derive the dependent variable Y as shown in figure 3 using the dependent variable selection algorithm, assuming that the rest of the variables X1, X2… Xn are independent.
3. Calculate the values of the coefficients b0, b1, b2…, and bn in the linear regression equation 13 from the training set using the least square multipliers
4. Use the multiple linear regression model equation deduced to calculate the missing values in the test set

4.1. MLRMUD can be used directly to estimate or predict the missing value if the row contains only one missing value.

4.2. If the row has more than one missing value, including the dependent variable, the independent attributes will be estimated using the Mean method and the dependent variable will be calculated by MLRMUD.

4.3. If the row has more than one missing value not including the dependent variable, then the second attribute found in the correlation test in figure 3 above, that is the second most repeated in the List, is calculated using MLRMUD after applying the Mean method on all the remaining missing values. The choice of the second attribute is due to having the second greatest number of correlated attributes.

IV. EXPERIEMENT AND RESULTS

The Iris dataset has been used to compare MLRMUD results to other recent approaches. The dataset is found in the UCI Machine Learning Repository [16]. It has four attributes and 150 records. The FPCM-SVRGA approach [6] has been tested using the Iris data set with various missing value ratios: 5%, 10%, 15% and 20%%. The missing values have been selected randomly for removal from the dataset. The accuracy of the results is evaluated using RMSE. Chhabra et al [7] used the Iris dataset where six approaches: Predictive Mean Matching, Multiple Random Forest Regression Imputation, Multiple Bayesian Regression Imputation, Multiple Linear Regression Using Non-Bayesian Imputation, Multiple Classification and Regression Tree (CART) and Multiple Linear Regression with Bootstrap Imputation have been tested using missing value ratio of 20%. They used a simple method to remove values

from the Iris dataset. The Iris dataset has petal width, petal length, sepal width and sepal length attributes. They removed 35 values from the petal width attribute, 30 values from the sepal width and the petal length attributes and 25 values from the sepal length attribute.

The Wine data set has also been used to compare MLRMUD approach to Bayesian Genetic Algorithm, BGA approach [17], with respect to RMSE. The BGA approach has been tested using various missing value ratios: 5%, 10%, 20%, 30%, 40%, 50% and 60%. No specific method was adopted to remove values from the dataset

A. *MLRMUD versus FPCM-SVRGA*

TABLE II. COMPARISON OF MLRMUD AND FPCM-SVRGA WITH RESPECT TO RMSE

Approach	Missing Value Ratio, %	RMSE
MLRMUD	5	0.1891
	10	0.3824
	15	0.3953
	20	0.2979
FPCM-SVRGA	5	0.565
	10	0.594
	15	0.6138
	20	0.6519

B. MLRMUD *vs Six Approaches [7]*

TABLE III. COMPARISON OF MLRMUD WITH SIX DIFFERENT APPROACHES WITH RESPECT TO MEAN STANDARD ERROR

S.NO	Method	Mean Standard Error
1	Predictive Mean Matching	0.10608496
2	Multiple random forest Regression Imputation	0.09765137
3	Multiple Bayesian Regression Imputation	0.09503033
4	Multiple linear regression using non-Bayesian imputation	0.11876531
5	Multiple classification and regression tree (CART)	0.10915661
6	Multiple linear regression with bootstrap imputation	0.11446101
7	MLRMUD	0.090175019

C. *MLRMUD vs BGA with respect to RMSE*

TABLE IV. COMPARISON OF MLRMUD WITH BGA ALGORITHM WITH RESPECT TO RMSE.

MV%	Approach	RMSE
5	BGA	1.6
10		1.72
20		1.81
30		4.1
40		5.2
50		7
60		10
5	MLRMUD	0.6
10		0.9242
20		1.2457
30		1.824
40		2.4256
50		4.245
60		5.545

D. *Dependent Variable Selection*

TABLE V. RESULTS OF THE PROPOSED DEPENDENT VARIABLE SELECTION ALGORITHM COMPARED TO THE ACTUAL DEPENDENT VARIABLE FOR DIFFERENT TRAINING SIZES.

DS#	Training Set%	Missing Value %#	Actual Unknown Dependent Variable	Proposed Algorithm	Comparing the results with the actual variable
1	100%	0%	Brain	Brain	T
	90%	5%		Brain	T
	80%	10%		Brain	T
	65%	15%		Brain	T
	50%	20%		Brain	T
	40%	25%		Brain Height Weight	T
	30%	30%		Height	F
	20%	35%		PIQ Brain Height Weight	T
	10%	40%		PIQ Brain Height Weight	T
2	100%	0%	X2	X2	T
	90%	5%		X2 X4	T
	80%	10%		X1 X2 X4 X5	T
	75%	15%		X2	T
	60%	20%		X2 X3	T
	50%	25%		X2 X3	T
	45%	30%		X2	T
	35%	35%		X2	T
	30%	40%		X2 X3	T
	16%	45%		X1 X2 X3 X4 X5	T
3	100%	0%	X1	X1 X2	T
	90%	5%		X1 X2	T
	80%	10%		X1 X2 X4	T
	70%	15%		X2	F
	60%	20%		X1 X2 X4	T
	50%	25%		X1 X4	T
	30%	30%		X1	T

	20%	35		X1	T
	15%	40		X4	F
4	100%	0%	Height	Height LeftArm RtArm LeftFoot RtFoot	T
	85%	5%		Height LeftArm	T
	70%	10%		Height	T
	60%	15%		Height LeftFoot RtFoot	T
	45%	20%		LeftFoot RtFoot	F
	35%	25%	Height	Height	T
	25%	30%		Height	T
	20%	35%		LeftArm	F
	5%	40%		LeftArm RtArm LeftHand RtHand	F
Accuracy of the dependent variable selection algorithm= number of truly detected cases / total cases					31/37=83%

V. CONCLUSION

In this work a new solving technique to the missing value problem with unknown, dependent variable (MV, UDV) is successfully introduced. The steps of the proposed approach are: splitting the data set into training and test sets, finding the dependent variable from the training set, creating a linear regression model from the training set and then using the model to predict missing values in the test set. The results show that the proposed model surpasses Fuzzy Possibilistic C Means Optimized with Support Vector Regression and Genetic Algorithm (FPCM-SVRGA) with respect to the RMSE. The minimum achieved RMSE is recorded as 0.1891. A comparison is also made with another six approaches which are Predictive Mean Matching, Multiple Random Forest Regression Imputation, Multiple Bayesian Regression Imputation, Multiple Linear Regression Using Non-Bayesian Imputation, Multiple Classification and Regression Tree (CART) and Multiple Linear Regression with Bootstrap Imputation with respect to the MSE for a missing value ratio of 20% where the proposed work outperforms these approaches with MSE of 0.090175019. MLRMUD is also compared with BGA with respect to the RMSE for missing value ratios of 5%, 10%, 20%, 30%, 40%, 50% and 60% % where it outperforms BGA for all the missing value ratios. The proposed-dependent variable selection algorithm achieves an accuracy of 83% over all the datasets examined.

REFERENCES

," Int. J. Information Technology and Management, Vol. 14, Nos. 2/3, 2015.

[1] Amir F. Atiya, Mohamed Aly, and Alexander G. Parlos, "Sparse basis selection: new results and application to adaptive prediction of video source traffic," IEEE Transactions on Neural Networks, Vol. 16, No. 5, pp. 1136-1146, September 2005.

[2] Jiří Kaiser, "Algorithm for Missing Values Imputation in Categorical Data with Use of Association Rules," Czech Technical University in Prague.

[3] Pier Alda Ferrari, Alessandro Barbiero, Giancarlo Manzi, "Handling Missing Data in Presence of Categorical Variables: a New Imputation Procedure, "March, 2011.

[4] Liqiang Pan, Jianzhong Li, "K-nearest neighbor based missing data estimation algorithm in wireless sensor network," doi:10.4236/ wsn.2010.

[5] Dempster AP, Laird NM, Rubin DB, "Maximum likelihood from incomplete data via the EM algorithm," JRSSB, 39:1–38, 1997.

[6] p.saravanan, p.sailakshmi, "Missing value imputation using fuzzy possibilistic c means optimized with support vector regression and genetic algorithm , " Journal of Theoretical and Applied Information Technology, February 2015.

[7] Geeta Chhabra, "A Comparison of Multiple Imputation Methods for Data with Missing Values," M Article in Indian Journal of Science and Technology, May 2017

[8] Amir Atiya and Yaser Abu-Mostafa, "An analog feedback associative memory," IEEE Trans. Neural Networks, Vol. 4, No. 1, pp. 117-126, January 1993.

[9] Mostafa Hassan, Amir Atiya, Neamat Gayar, Raafat El-Fouly, "Novel ensemble techniques for regression with missing data," New Mathematics and Natural Computation 5 (03), 635-652, 2009.

[10] Tawfik A. Mohamed, Neamat El Gayar, Amir F. Atiya "Forward and backward forecasting ensembles for the estimation of time series missing data," IAPR Workshop on Artificial Neural Networks in Pattern Recognition, 93-104, 2014.

[11] Robert Andrawis and Amir F. Atiya, "A new Bayesian formulation for Holt's exponential smoothing , " Journal of Forecasting, Vol. 28, No. 3, pp. 218-234, April 2009.

[12] Dhwani Jayant Raval , " Big data analytics for finding missing values using linear regression technique for numeric datasets , " August 2015 .

[13] Mr.M.B.Shelke,Mr.K.B.Badade," Processing of Incomplete Data Sets: Prediction of Missing Values by using Multiple Regression," M.B.Shelke, et al International Journal of Computer and Electronics Research Vol.2, No. 5, October 2013.

[14] Hailin , "Incomplete Data Recovery Using Linear Regression,"Applied mechanics and material Vols.571-572, 2014.

[15] Yasser Beyad, Marcel Maeder ", Multivariate linear regression with missing values , "Analytica Chimica Acta xxx (2013) xxx–xxx.

[16] UCI Machine Learning Repository,"https://archive.ics.uci.edu /ml/datasets.html".

[17] R. Devi Priya and S. Kuppuswami ," A novel approach for imputation missing continuous attribute values in databases using genetic algorithm