INTERNATIONAL JOURNAL ON HUMAN COMPUTING STUDIES

www.journalsresearchparks.org/index.php/IJHCS e-ISSN: 2615-8159|p-ISSN: 2615-1898

Volume: 02 Issue: 01 | Apr-May 2020

Multivariate imputation for missing data handling a case study on small and large data sets

Yagyanath Rimal

Pokhara University, Nepal

EMAIL: rimal.yagya@pu.edu.np

Abstract - *Abscent of records generally termed as missing data which should be treated properly before analysis procedes in data analysis. There were many researchers who undoubtedly mislead their research findings without proper treatment of missing data, therefore this review research try to explain the best ways of missing data handling using r programming. Generally, many researchers apply mean and median imputation but this sometimes creates bios in many situations, therefore, the researcher tries to explain some basic association among other research variables with treating missing data using r programming. The imputation process suggests five alternatives be replaced for missing data values were generated automatically and substituted easily at the process of data cleaning and data preparation. Here researcher explains two sample data for missing treatment and explains many ways for graphical interpretation of them. The first data set with 12 observation describes the easiest way of missing replacement and the second vehicle failure data from internet of 1624 records, whose missing pattern were calculated and replaced with to the respective data sets before analysis.*

Key Words: Not Available, Multivariate Imputation via Chained Equations, Visualization and Imputation of Missing Values.

1.INTRODUCTION

Not available data is one of the main problems for data scientists before analyzing it, in many cases, the missing data must be analyzed after proper correcting in the data analysis process incomplete cases. According to Julian Williams, there is no real implementation, if the data that has more than 5 percent of the data sets with NA (Loo, 2013). There have been many situations sometimes considered by the researcher, some people absent at some time, at a certain point the respondent will not provide the appropriate answer for many reasons and, sometimes interviewees have not answered insensitive questions also because they create missing data (NZ, 2015) unknowingly. This situation leads to prejudices for scientific data researchers to manage them without significant treatment first. The data may be missing due to many reasons. Horton and Kleinman (2007) in a review of medical research (2001, 49) reported that about 94 percent of analyzes use removing the list to eliminate complete observations (Gary King, 1998). This process leads to a loss of valuable information further creative bias in research. There are many ways to correct them, but none of them will be free from partiality and adverse effects. Sometimes, data scientists will apply simply ignoring the tuple, but this leads to the elimination of other precise data followed which, ultimately, to the loss of data (Géron, 2017). At some point, linear regression is calculated, but this process does not exclude the data that is not appropriate to complete the respondents too. If the investigator enters NULL or N / A, the sample size is also sometimes reduced largely. The most common way to correct is to replace with a median mean replacement (Verzani, 2012) are always adequate to replace with a median in the datasets of larger sample data. Therefore, the data generation algorithm always suggests some values after having machine learning decision trees for grouping of replacements of data. In spss there are tools to restore missing data, but they are not always accurate for all datasets, however, these procedures have several processes of replacement and insertion (Roy, 2019). The missing at random pattern is corrected on dependent variables with calculating of probability for recommending for replacement values. Likewise, the non-random pattern primarily dependent on variables such as low-income association are less likely to respond to others, therefore our conclusions for replacement will be not exactly matched absolutely. The not random pattern is the worst case where, there have no procedures recommended by the data specialist (Sauro, 2015). It is applicable only

INTERNATIONAL JOURNAL ON HUMAN COMPUTING STUDIES
www.journalsresearchparks.org/index.php/IJHCS e-ISSN: 2615-8159|p-ISSN: 2615-1898
Volume: 02 Issue: 01 | Apr-May 2020

when there is a large sample enough for substantial loss of statistical power. However, if there are missing

Sn	y	z	q	complete
1	12	11	4	1
2	13	13	5	1
3	14	14	6	1
4	15	15	34	1
5	17	16	5	1
6	NA	17	6	0
7	12	18	3	1

values in a random model that causes the involuntary deletion of observations. At some point the recontact the participants , ask them to complete the missing values are not possible, therefore the imputation is the best alternative for replacing the missing values. The random arbitrary replacement is not the researcher preferred course of action, but often a missing value can be inferred wrong (Oehlert, 2010). The average value missing value substitution is not always recommended because it can artificially reduce data variability, but in some cases, this makes sense too (Howard J, 2018). Allocation of the common point for the most commonly chosen midpoint or value. This is a little more structured than a hypothesis, but it is still among the riskiest options (Alma, 2017). Therefore, regression replacement analysis to estimate is quite stable regression automatically predicts the lost values. The multiple imputations, the most sophisticated approach is to further convey the idea of regression and take advantage of correlations between correlations of responses for missing data and therefore average simulated data sets incorporating random errors in the forecasts (Him, 2010) recommendation. The best measurable bundles like SPSS have various attribution work (Aizaz Chaudhry, 2018) similarly but information resembles a medicinal to overlook doesn't cause properly. The KNN distribution strategy of missing estimations utilizes the closeness of two resolved fact that by utilizing the K closest neighbor technique for both subjective and quantitative qualities with additionally missing qualities that can be prepared best on for huge databases. Therefore R bolsters MICE, Amelia, missForest, Hmisc bundles are the best tools for missing information. The MICE has utilized data clients (Vidhya, 2016) linear regression to anticipate consistent missing qualities were calculated for lost absolute qualities (Pasteels, 2013). The amelia package additionally plays out different attributions since it is empowered with a launcher to ascribe numerous factors (Heeringa, 2019). Typical multivariate dissemination (MVN) utilizes means and covariance to abridge information that is overseen haphazardly. Also, miss forest backwoods is a usage of the arbitrary woodland calculation that uses a nonparametric ascription (zwiak, 2018).The nonparametric strategy doesn't cause express suppositions about practical endeavors to gauge f so it to can be so near the information focuses without seeming unrealistic. Produces a gauge of the OOB ascription blunder. Essentially, hmisc is a multipurpose package helpful for information investigation of missing qualities, propelled table creation, adjustment and model analysis (Buuren, 2019). Also, the mi (Various ascriptions) bundles give a few highlights to managing missing qualities to rough missing qualities for the prescient strategy for media coordinating (Su, 2011).

Imputing Using the Mean/ Mode Suppose that we have a simple vector of marks of computer science subjects are 95, 88, 85, NA, 75, 70, 78, NA, 70 and 68 data sets. The data in the given frame indicates two missing values represented with NA i.e not applicable or missing, whose statistics calculation will be wrong without replacing NA with numeric values, due to un-imputed values in fourth and eight indexes, therefore missing values will be addressed properly using im-putting. The variable [is.na ()] function always extract the missing values NA twice similarly if we inverse the function variable [!is .na()] gives all the data except NA in the vector elements. If the calculated after excluding NA this may be wrong in many times where there were lesser items of elements , however, we

INTERNATIONAL JOURNAL ON HUMAN COMPUTING STUDIES
www.journalsresearchparks.org/index.php/IJHCS e-ISSN: 2615-8159|p-ISSN: 2615-1898
Volume: 02 Issue: 01 | Apr-May 2020

could calculate statistical values. Therefore we need mean or median imputation of data is quite safety procedures for overcome such problems. The original data is to stored in another variable then variable[!is.na (variable)]=mean(variable [!is.na(variable))] then the data frame will be as 95.0, 88.0, 85.0,**78.2**,75.0,70.0,78.0, **78.2**,70.0, 68.0 here mean 78.2 will be replaced with NA. Similarly median imputation produces the 95,88, 85,**77**, 75,70, 78,**77**,70,68 completes its NA without losing its numbers of items in the data frame.

2. SIMPLE LINEAR REGRESSION IMPUTATION FOR SMALL DATASETS

From the above table the variable y, z and q were data values the y has three missing at 6, 9 and 10 position where z and q variables were not any missing data were taken as sample data. First the researcher needs to calculate which records have missing items. The correlation of varialbes is calculated and imputed in those NA places considering all rest items will not change when complete case is zero. The which function gives the index of NA values. The which (is.na(data)) gives 6 ,9 and 10 position. If we inverse which(!is .na(data)) gives all numeric values except NA items of y variable. The above concept is apply using r programming as

```
> x=1:12
> y=c(12,13,14,15,17,NA,12,43,NA,NA,18,17)
> z=c(12,13,14,15,16,17,18,13,10,11,19,20)
> q=c(4,5,6,34,5,6,3,2,4,5,6,7)
> data=data.frame(x,y,z,q)
> data #Displays the output like above table.
> cor(data)
          x         y         z         q
x  1.0000000 NA 0.34511839 -0.18747823
y         NA     1        NA         NA
z  0.3451184 NA 1.00000000  0.09106705
q -0.1874782 NA 0.09106705  1.00000000
```

When calculating correlation the NA values do not calculated due to missing values in y column.

```
> cor (data, use="complete.obs")
          x         y         z         q
x  1.0000000  0.3552472  0.833268458 -
0.143288779
y  0.3552472  1.0000000 -0.186104020 -
0.192336337
z  0.8332685 -0.1861040  1.000000000
0.004399981
q -0.1432888 -0.1923363  0.004399981
1.000000000
```

The use equal to complete observation calculate without NA correlation. The synum describe the correlation between large vairalbesindicating with dot and Multiplication symbol indicates highest correlation between x and y varialbes.

```
> int=function (t){
+   x=dim (length(t))
+   x[ which (! is.na (t))]=1
+   x [ which ( is.na (t ) ) ] =0
+   return (x) }
```

The function int return 1 if there is non missing where 0 for missing values with return tpe and arguments as t.

```
>  data $ comp=int (data $y)
>  data
   x  y  z  q comp
1   1 12 12  4    1
2   2 13 13  5    1
……..
12 12 17 20  7    1
```

The new column added tith missing or without missing indication in data

```
> lm (y~ x, data = data)
Lm (formula = y ~ x, data = data)
Coefficients:
(Intercept)          x
   12.6820      0.8842
```

The linear regression is calculated with the help of y with x where data = data and produced intercept and slope. .

```
> for (i in 1:nrow (data ) ){
+   if (data $comp [i] == 0) {
+     data $y[i]=12.68+0.8842*data$x[i]   }}
```

The for loop is imputing the data cells when comp is equal to only 0 with missing values on the basic of its intercept and slope.

```
> data
   x      y  z  q comp
1   1 12.0000 12  4    1
2   2 13.0000 13  5    1
……..
12 12 17.0000 20  7    1
```

INTERNATIONAL JOURNAL ON HUMAN COMPUTING STUDIES
www.journalsresearchparks.org/index.php/IJHCS e-ISSN: 2615-8159|p-ISSN: 2615-1898
Volume: 02 Issue: 01 | Apr-May 2020

The 6, 9,10 positional values were imputed in y variables with 17.98,20.63 and 21.52 values become non-missing values. So that we could easily calculate summary with min, max, 1st quartile and median, mode of statistics tool after linear regression imputation.

> summary(data)

The mean is always chosen than the other data in a column with normal distribution otherwise median is chosen with the data having the right and left skewness of the pattern. Some time mode is chosen when the data are in loss of suicidal cases of a time interval. In some time we may design a separate model for calculating the missing model while design model the missing data are set for test purpose and impute the missing value then apply normal process of test and training test data sets of machine learning procedures.

3. USING R PROGRAMMING FOR LARGER DATASETS

Here researcher uses vehicle failure data sets for treatment of missing data which has five variables and 1624 observational records. Sometimes it is a very bad idea to replace mean median replacement for missing data case therefore here multivariate imputation chain equation (MICE) and visualization and imputation of missing values VIM package was used for the treatment of NAs and lattice is a powerful and elegant high-level data visualization system inspired by trellis graphics. It is designed with an emphasis on multivariate data, and in particular, allows easy conditioning to produce plots. After loading system library the data structure

> str(data)

classes tbl df tbl and data.frame 1624 obs. of 5 variables:

The str function displays the data structure of vehicle failure vehicle, mileage hours and cost and state records of 1624 observations were analyzed.

> summary(data) NA's :5 NA's :31 NA's :30

The summary commands display statistical summary including NAs fo each column but the categorical variable state should treat with factor

> data$State=factor(data$State)

> data.frame(data)

Vehicle	Mileage	lh	lc	State
1	863	1.1	66.3	MS
2	4644	2.4	233	CA
3	16330	4.2	325.1	WI
4	13	1	66.64	OR
5	22537	4.5	328.7	MS
6	40931	3.1	205.3	CA
7	34762	0.7	49.17	WI
8	11051	2.9	208.8	OR
1624	24879	3.5	260.3	WI

Vehicle Mileage lh lc State
1 1 863 1.1 66.30 MS
2 2 4644 2.4 233.03 CA
……
199 199 4406 5.6 421.28 OR

After treating state variable summar command demonstrate its missing values, the milage with 5 NAS, labour hour has 31 missing and Lcost has 30 and state has 11 missing NAS presents others were no missing data.

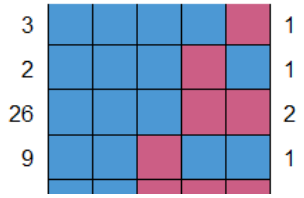
> p=function(x){sum(is.na(x))/length(x)*100}
> apply(data,2,p)#1 for row 2 for column

Vehicle Mileage lh lc State
0.0 0.3078818 1.9088670 1.8472906 0.6773399

The percentage of missing values is calculated with total by its length demonstrate vehicle has no missing data similarly millage has .30 percent, 19 percent of labor hours, 18 percent and the state has .67 percent respectively.

> md.pattern(data) # 1577 records have not missed a single value out of 1624.

Vehicle Mileage State lc lh
1577 1 1 1 1 1 0
3 1 1 1 1 0 1
2 1 1 1 0 1 1
26 1 1 1 0 0 2
9 1 1 0 1 1 1
2 1 1 0 0 0 3
5 1 0 1 1 1 1
0 5 11 30 31 77



INTERNATIONAL JOURNAL ON HUMAN COMPUTING STUDIES
www.journalsresearchparks.org/index.php/IJHCS e-ISSN: 2615-8159|p-ISSN: 2615-1898
Volume: 02 Issue: 01 | Apr-May 2020

Out of 1624 total records, there were 77 records missing, the milage has 5, the state has 11 lc and ln 31 and 30 not available records. Similarly, the 2 rows have a maximum 3 records of missing and the vehicle has no missing records.there are 26 records has two missing labor cost and labor hours.

```
> md.pairs(data)# rm overfed vs  missing
$rr
        Vehicle Mileage   lh   lc State
Vehicle   1624    1619 1593 1594  1613
Mileage   1619    1619 1588 1589  1608
lh        1593    1588 1593 1591  1584
lc        1594    1589 1591 1594  1585
State     1613    1608 1584 1585  1613
```

The above data demonstrate observed and missing

```
$rm
        Vehicle Mileage lh lc State
Vehicle       0       5 31 30    11
Mileage       0       0 31 30    11
lh            0       5  0  2     9
lc            0       5  3  0     9
State         0       5 29 28     0
```

Demonstrate observed and missing of each variables

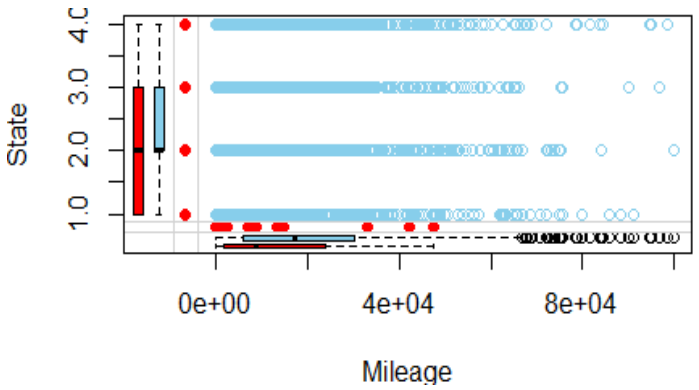
```
$mr
        Vehicle Mileage lh lc State
Vehicle       0       0  0  0     0
Mileage       5       0  5  5     5
lh           31      31  0  3    29
lc           30      30  2  0    28
State        11      11  9  9     0
```

Missing vs observed

```
$mm
        Vehicle Mileage lh lc State
Vehicle       0       0  0  0     0
Mileage       0       5  0  0     0
lh            0       0 31 28     2
lc            0       0 28 30     2
State         0       0  2  2    11
```

Demonstrate missing vs missing records.

```
> marginplot (data[,c('Mileage','State')])
```



The marginal plot was calculated mileage and state records where red dots were missing of state and blue were mileage records.

```
> impute=mice (data [ , 2:4], m=3, seed=123)
```

The imputation is carried out by mice package with second to the fourth column and imputed options were specified by 3 but there were 5 imputations with same seed 123.

```
 iter imp variable
  1   1  Mileage  lh  lc
  .............
  4   3  Mileage  lh  lc
  5   3  Mileage  lh  lc
```

The iteration and imputation carried out in every three items of five items were calculated.

```
> print(impute)
Number of multiple imputations:  3
Imputation methods:
Mileage     lh     lc
 "pmm"  "pmm"  "pmm"
PredictorMatrix:
        Mileage lh lc
Mileage       0  1  1
lh            1  0  1
lc            1  1  0
```

the predictive mean method is used for numeric variable and the state has polreg methods multinomial logistic regression were used.

```
> impute$imp$Mileage
        1    2    3
1564 12455 21610 47357
1565 17636 42324 18506
1566 35021 17261 26348
1622  7289 13823 24130
1623     4 67903 16963
```

INTERNATIONAL JOURNAL ON HUMAN COMPUTING STUDIES
www.journalsresearchparks.org/index.php/IJHCS e-ISSN: 2615-8159|p-ISSN: 2615-1898
Volume: 02 Issue: 01 | Apr-May 2020

This demonstrates three imputed value of mileage optional data for imputation but whose data were again verified as.

```
> data[1564,]
1   1564     NA   5.1  371. OR
> data[1566,]
1   1566     NA   2.2  154. CA
> data[1622,]
1   1622     NA   4.5  318. CA
> summary(data$Mileage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      1    5778   17000   20583   30100   99983       5
```

From the above records the mean, median values were imputed in either of three values with mean 20583 values whose labor hours are only 4.5 is not justified the second imputation with 17261 is taken best.

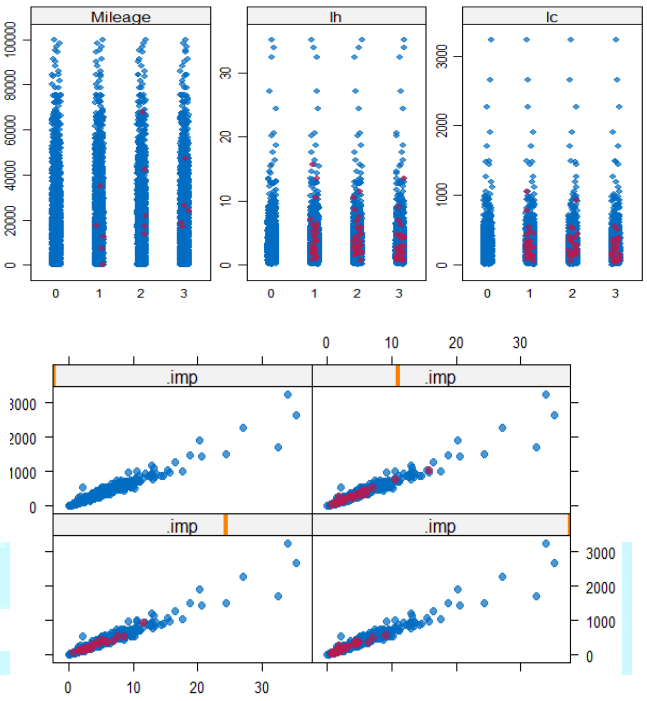
```
> new=complete(impute,2)
> new
   Mileage  lh     lc
1      863  1.1   66.30
2     4644  2.4  233.03
3    16330  4.2  325.08
```

```
330    6074  4.3  282.75
331   13540  6.2  492.17
332   10012  4.9  419.31
333   10923  3.1  231.98
new[1564,]
    Mileage  lh     lc
1564   21610 5.1 370.77
> new[1566,]
    Mileage  lh     lc
1566   17261 2.2 154.49
> new[1622,]
    Mileage  lh    lc
1622   13823 4.5 318.1
> stripplot (impute,pch=20,cex=1.2)
```

This plot demonstrates missing values and observed value distribution in each column data. The second box plot demonstrates first-second and third imputed values in whole data sets which does not demonstrate unusual patterns on original data sets after imputation.

```
> xyplot (impute,lc~lh|.imp,pch=20,cex=1.4)
```

This plot demonstrates a relationship between labor hours and labor const matches first, second and third imputation describes patterns after imputation.



4. CONCLUSION

From the above procedures, the researcher could easily impute missing data with imputing mean and average data. The AI preparing model with a dataset that has a ton of missing qualities can definitely affect the AI model's quality. The k closest neighbors is a calculation that is utilized for straightforward order just. The numerous attributions are obviously superior to a solitary ascription as it gauges the vulnerability of the missing qualities in a superior manner. In spite of the fact that there is no ideal method to appropriately address missing qualities before investigation. Every system can perform better for certain datasets and missing information types. There are some set standards to choose which system to use for specific kinds of missing qualities which model works best for your dataset.

REFERENCES

[1] Aizaz Chaudhry, W. L. (2018). A Method for Improving Imputation and Prediction Accuracy of Highly Seasonal Univariate Data with Large Periods of Missingness. Academic Editor: Simone Morosi.

[2] Alma, P. E.-F. (2017). © 2017 Pedersen et al. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at https://www.dovepress. com/terms.php and incorporate the Creative Commons Attribution – Non-Commercial (unported, v3.0. Clinical Epidemiology downloaded from https://www. Dove press. com / by 128.

[3] Buuren, S. v. (2019). Package 'mice'December 13, 2019.

[4] Gary King, J. H. (1998). Listwise Deletion is Evil: What to Do About Missing Data in Political Science.

[5] Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn & TensorFlow. Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurélien Géron. All rights reserved. Printed in the United States of America.

[6] He, Y. (2010). Missing Data Analysis Using Multiple Imputation: Getting to the Heart of the Matter.

[7] Heeringa, P. B. (2019). Multiple Imputation of Missing Data Using SAS.

[8] Howard J, S. (2018). Experimental Design andAnalysis.

[9] Loo, E. d. (2013). An introduction to data cleaning with R. Statistics Netherlands.

[10] The Hague/Heerlen 2013.Reproduction is permitted, provided Statistics Netherlands is quoted as the source.

[11] NZ, S. (2015). A guide to good survey design (4th ed). Available from. Statistics NZ Information Centre: info@stats.govt.nz.

[12] Oehlert, G. W. (2010). A First Course InDesign and Analysis of Experiments. Design-Expert is a registered trademark of Stat-Ease, Inc.Library of Congress Cataloging-in-Publication Data.

[13] Pasteels, J.-M. (2013). 1 Review of best practice methodologies for imputing and harmonizing data in cross-country datasets.

[14] Roy, B. (2019). Missing Data Imputation Techniques. Data Science | Machine Learning | Deep Learning | Artificial Intelligence.

[15] Sauro, J. (2015). 7 Ways to Handle Missing Data . W AY S T O H A N D L E M I S S I N G DATA 2, 2015.

[16] Su, Y.-S. (2011). Multiple Imputation with Diagnostics (mi) in R.

[17] Verzani, J. (2012). simpleR{UsingR forIntroductoryStatistics. JohnVerzani(verzani@ math. csi. cuny.edu), 2001-2.All rights reserved.

[18] Vidhya, A. (2016). Tutorial on 5 Powerful R Packages used for imputing missing values.

[19] ´Zwick, K. J. (2018). What to do when incomplete subjects information is available?Tutorial on Dealing with Missing Data.