Imputation Techniques To Handle Missing Data

Student Name: Dhruv Shah
Enrollment ID: **202103017**
Summer Research Internship (SRI) Report
SRI Mode: ***On Campus***
Dhirubhai Ambani Institute of ICT (DA-IICT)
Gandhinagar, India
202103017 [at] daiict.ac.in

Mentor's Name: Prof. Tathagata Bandyopadhyay
Dhirubhai Ambani Institute of ICT (DA-IICT)
Near Indroda Circle
Gandhinagar 382007, India
tathagata_b [at] daiict.ac.in

***Abstract*—On a regular basis we encounter datasets with missing values, this creates problems to analyse and draw meaningful insights from such datasets. A common technique to handle the missing values is to use mean imputation, but this technique has many disadvantages such as underestimation of standard errors and introducing bias in the data. To tackle this problem, popular techniques such as the Expectation Maximization Algorithm and Multiple Imputation are used. In this report, we study these algorithms from the existing literature. Then we try to apply these algorithms to analyse the partially contested elections in the Indian Electoral System.**

***Index Terms*—missingness, imputation techniques, maximum likelihood, expectation maximization, multiple imputation.**

I. Introduction

In today's data driven world, almost everyone encounters the need to work with data on a regular basis. Proper analytics of data is required in almost all fields ranging from healthcare and pharmaceuticals to defence and security. Many such datasets contain missing values which make the analysis of these datasets even harder. We begin by defining a missing value and types of missingness that occur in datasets.

A. What is a missing value?

Any entry in a dataset which is Not Available (NA) in a dataset can be termed as a missing value. Missing values include numbers that have been grouped, aggregated, rounded, censored or truncated resulting in partial loss of information [1]. Missing values that occur in any data collection procedure such as surveys are also termed as non-responses.

B. Types of Non-Responses

There are two types of non-response that can occur.

- Unit Non-Response: This occurs when an entire data collection procedure fails. For example, a person refuses to participate in a survey or is absent during the entire survey.
- Item Non-Response: This occurs when partial data is available. For example, a person participates in the survey but does not respond to certain individual items.

II. Distributions of Missingness

For any dataset we can define indicator variables R that identify what is known and what is missing. For example, R can be a binary variable taking the value $R = 0$ if a value is missing and $R = 1$ if a value is present. The varialbe R is referred to as *missingness* [1].
In a real world dataset it becomes impossible to describe accurately the reasons for missingness, the distribution of R is regarded as a mathematical device to describe the pattern of missingness, helping to capture the relationship between missingness and the missing values.
We classify the types of distributions for R according to the nature of their relationship with the data. Let us denote the complete data as Y_{com} and partition it as $Y_{com} = (Y_{obs}, Y_{miss})$, where Y_{obs} and Y_{miss} are the observed and missing parts of the data respectively.
The distributions of Missingness are classified in the following three categories.

A. Missing At Random

If the distribution of missingness does not depend on Y_{miss}, then such missingness is defined as Missing At Random (MAR) [1]. In terms of conditional probability, it can be written as,

$$P(R|Y_{com}) = P(R|Y_{obs}) \tag{1}$$

B. Missing Completely At Random

This is a special case of MAR, when the distribution does not depend on Y_{obs}. Therefore, it can be written as,

$$P(R|Y_{com}) = P(R) \tag{2}$$

	X	Y	mcar	mar	mnar
1	150.26501	156.21602	156.2160	156.2160	156.2160
2	121.91686	113.49354	NA	NA	NA
3	127.43993	182.03050	NA	NA	182.0305
4	162.34455	144.56022	NA	144.5602	144.5602
5	134.88521	133.65812	133.6581	NA	NA
6	87.76901	93.36234	NA	NA	NA
7	97.32444	111.14878	111.1488	NA	NA
8	120.33984	116.47986	116.4799	NA	NA
9	149.73500	100.00708	NA	100.0071	NA
10	192.89732	164.64205	NA	164.6420	164.6420
11	138.06376	146.08518	NA	NA	146.0852
12	107.25710	107.01012	NA	NA	NA
13	109.85439	88.82083	NA	NA	NA
14	119.82241	117.23248	NA	NA	NA
15	127.42097	109.19772	NA	NA	NA
16	113.08975	118.50693	NA	NA	NA
17	146.57297	114.70680	NA	114.7068	NA
18	100.96469	109.14738	NA	NA	NA
19	131.96484	137.51951	137.5195	NA	NA
20	96.59786	98.05774	NA	NA	NA

Fig. 1. A snap shot of the values of blood pressure for 20 patients out of 50. Columns X and Y show the complete data, with missing values imposed by the three different methods on Y

C. Missing Not At Random

If the distribution depends on Y_{miss} and the Equation 1 is violated then the missing data is called Missing Not At Random (MNAR).

Now let us illustrate all the three types of missingness using an example.
We will simulate data for the systolic blood pressure of 50 individuals who come for check-up at a clinic for the month of January(X) and February(Y). Let the data for both the months be drawn from a bivariate normal population with mean $\mu_x = \mu_y = 125$, standard deviations $\sigma_x = \sigma_y = 25$ and covariance $\rho = 0.60$.
Now, to impose the missing values on the column Y using the MCAR method, we randomly select data of 10 people from X and take their readings of Y. Rest all are kept as NA.
For MAR, we select the February readings, i.e. the value of Y for those people whose blood pressure reading for January is greater than 140, i.e. $X > 140$.
For the MNAR mechanism, we select only those individuals whose reading for the month of February (values of Y) are greater than 140. This could have happened in a case when all the 50 patients had returned for the checkup in February but the staff person in-charge decided only to keep those entries whose value exceed 140.

We will now apply imputation techniques to handle the missing data created in this example. Primarily we will study the Expectation Maximization (EM) Algorithm and the Multiple Imputation (MI) technique.

III. Imputation Techniques

A. Expectation Maximization (EM) Algorithm

The Expectation Maximization Algorithm is applied to obtain the maximum likelihood estimates of the parameters (in the bivariate normal case it is μ and Σ) when some of the data is *missing*. This algorithm is applied under the general assumption that the underlying data is Missing At Random (MAR), i.e. the unobserved data was never intended to be observed in the first place.
Mathematically, the likelihood function can be written as

$$l\left(\theta; Y_{obs}\right) = \log f\left(Y_{obs}; \theta\right) = \log \int f\left(Y_{obs}, Y_{miss}; \theta\right) dY_{miss}$$

here θ is the unknown parameter and $f(Y;\theta)$ is the probability density function of $Y = (Y_{obs}, Y_{miss})$.

As we do not have the complete data, the EM Algorithm maximizes the expected complete data likelihood function. The details of the algorithms are described as following.

- E-Step or Expectation Step: This step computes the expected value of the log-likelihood function given the observed data. i.e. it calculates the conditional expectation using the parameter estimate of the current iteration.
  Let the current parameter be $\theta^{(t)}$, then

$$Q\left(\theta, \theta^{(t)}\right) = \mathbb{E}\left[l(Y;\theta) \mid Y_{obs}; \theta^{(t)}\right]$$

$$= \int l(Y;\theta) f\left(Y_{miss} \mid Y_{obs}; \theta^{(t)}\right) dY_{miss}$$

- M-Step or Maximization Step: This step computes the maximum likelihood estimate over the expectation computed in E-Step. Therefore, we get

$$\theta_{\text{new}} := \max_{\theta} Q\left(\theta; \theta^{(\text{t})}\right)$$

- Now, update $\theta^{(t)} = \theta_{new}$

We implement this algorithm for the example illustrated in Section-2 and reproduce the results obtained in [1]. The R-code for the implementation can be found here.

B. Multiple Imputation

Multiple Imputation is a technique to handle missing data problems which was developed by Rubin [2].

Even when an algorithms such as the EM Algorithm or the Stochastic Regression preserve the joint and the marginal distributions of data, they fail to take into account the fact that in almost all the cases we can never recover the missing data. Therefore

to account for this missing data uncertainty, and Multiple Imputation is one of the solutions to the problem.
The general outline for this procedure is as follows : Impute each of the missing values in the dataset with M values, generating a total of M imputed datasets. Each time the missing data is imputed using an imputation model, say for example, a regression model applied on the observed data and then predicting it for the missing data. We refer to the book [3] for the general MI Procedure:

- For $i = 1, 2, ..M$, impute each missing value using an imputation model generating a total of M 'completed' datasets.
- Fit the substansive model to i^{th} imputed dataset (now this does not have any missing values), $i = 1, 2, ..M$. Therefore, we obtain M estimates of the parameters say $\hat{\beta}_i$ and their variances $Var(\hat{\beta}_i)$
- Now, we combine these results using Rubin's rules to obtain point estimates.

Rubin's rules are stated as follows: The Multiple Imputation Estimate of β is as follows:

$$\hat{\beta}_{MI} = \frac{1}{M} \sum_{i=1}^{M} \hat{\beta}_i,$$

with variance

$$\widehat{V}_{MI} = \widehat{W} + \left(1 + \frac{1}{M}\right) \widehat{B},$$

where

$$\widehat{W} = \frac{1}{M} \sum_{i=1}^{M} \hat{\sigma}_i^2 \quad \text{and} \quad \widehat{B} = \frac{1}{M-1} \sum_{i=1}^{M} \left(\hat{\beta}_i - \hat{\beta}_{MI}\right)^2$$

IV. Partially Contested Elections As A Missing Data Problem

We take forward our understanding of the missing data techniques to the analysis of partially contested multi-party elections. Honaker et. al. [4], first developed this method to treat the problem as one of missing data. They consider the vote share of non-contesting parties in a particular district Missing at Random and apply the Multiple Imputation technique as described above.

V. Analysing the Partially Contested Indian Lok Sabha Elections

In the electoral system of India, we have multiple parties contesting an election as opposed to the bi-party system in the US. The Multi-party system in India is suitable because of the vast diversity seen in different regions of the country. Many parties form coalitions in order to maximize their chances of getting elected to form a government.
Here, we consider districts where coalitions are formed as partially contested elections and apply the multiple imputation technique as described by [4]. The goal of our analysis would be primarily to explain the effective vote share of the parties if there were no coalition between them.

In our implementation of the above problem, we use the 2014 Lok Sabha election data for the Punjab constituency, where the Bhartiya Janta Party (BJP) and the Shiromani Akali Dal (SAD) had formed a coalition to contest the election. We preprocess the data and implement the MI using the Amelia Package of R.

VI. Future Scope

We will implement the Multiple Imputation technique to analyse the partially contested elections due to coalitions formed in the states of Punjab and Uttar Pradesh for the 2014 Lok Sabha Elections.
We can introduce covariates such as incumbency advantage, economic diversity of the voters in the constituency, demographic indicators of the constituency such as the percentage of women voters of a political party and many more such indicators which might influence the vote share of a political party.

VII. Learning Outcomes

A. Technical Learnings

- **Missing Data Problem**: Learnt about importance of handling missing data to make efficient analysis of the data in hand.
- **Missing Data Imputation Techniques**: Developed an understanding of the state-of-art techniques to handle missing data.
- **R Programming**: Got familiar with the R-Programming Language to code the missing data techniques learnt.

B. Non-Technical Learnings

- **Time Management**: During the summer, I also prepared for the GRE along with this research internship. This honed my time management skills since I needed to allocate proper time for both.
- **Communication and Presentation**: Effective communication between me and my mentor was a key aspect which helped me learn a lot of new things in a short period of time.

ACKNOWLEDGEMENTS

I would express my immense gratitude to Prof. Tathagata Bandyopadhyay for providing me this opportunity to learn this topic under his guidance. I would also like to thank my friend and senior Nisarg Suthar for introducing me to this topic and motivating me to explore this area from a research perspective. Also, I would like to thank my parents and my younger brother for always being there for me.

REFERENCES

[1] Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002 Jun;7(2):147-77. PMID: 12090408.

[2] Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons Inc., New York. http://dx.doi.org/10.1002/9780470316696

[3] Carpenter, James R., and Michael G. Kenward. Multiple Imputation and Its Application. 1st ed, John Wiley & Sons, 2013

[4] Honaker J, Katz JN, King G. A Fast, Easy, and Efficient Estimator for Multiparty Electoral Data. Political Analysis. 2002;10(1):84-100. doi:10.1093/pan/10.1.84