The analysis of missing data for Bone Mineral Density

Katleho Nyoni

SUMMARY

The aim of this analysis is to create and predict the missing data in the Bone Mineral Density(BMD) dataset so that an Analyst can be able to perform classical statistical & mathematical techniques to answer research/business problems. I imputed a 1000 values using 20 iterations with a seed of 10 by using Multiple Imputation by Chained Equations(MICE).

1 INTRODUCTION

The problem of incomplete data across different variables has been a problem in many fields for decades. Classical statistical techniques cannot handle missing values and often exclude missing data, this is known as the deletion method when it comes to handling data. This makes our sample size smaller, resulting and advocating for inconsistent and inefficient inferences. Multiple Imputation(MI) is one of the recently developed methods to deal with missing data and one of the most efficient tool we have. Although the aim of this analysis isn't to review the methodology of MI, let us briefly go through it so the reader without knowledge of this concept doesn't get lost. Data might be missing for a number of reasons ranging from Missing At Random(MAR), Missing Completely At Random(MCAR) or Missing Not At Random(MNAR). It is assumed that the process that makes data to be missing is random, and this process is called the Missing Data Mechanism(MDM). Traditionally, methods that were used to deal with missing data was complete-case methods or single-imputation(SI) methods. It was until recently when (Rubin,1989) developed the now widely used Multiple-Imputation(MI). MI methods include Multivariate Normal MI, Bayesian MI and Multiple Imputation by Chained Equations(MICE)/Sequential Regression MI(SRMI)/Joint Conditional Specification among others. This analysis focuses on the latter for flexibility and robustness reasons.

1.1 DATASET

We will be using the BMD dataset obtained from Kaggle. This dataset is made of 9 variables of the form

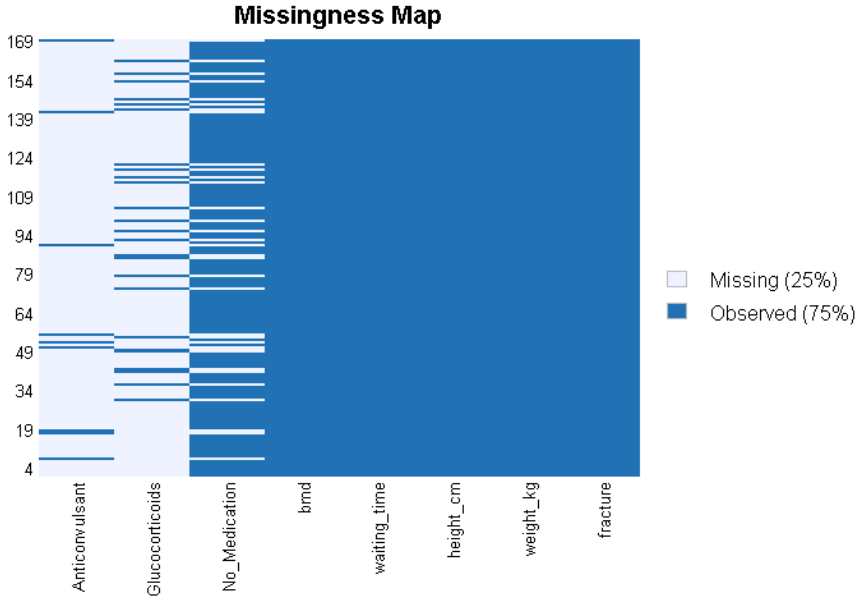
1. id - id of participants
2. age - Continuous
3. sex – Two level factor : Male, Female
4. fracture - Two level factor: Fracture, No-Fracture

5. weight_kg - Continuous
6. height_cm - Continuous
7. medication – Three level factor: No Medication, Glucocorticoids, Anticonvulsant
8. waiting_time - Continuous
9. bmd - Proportions

After restructuring the dataset using simple techniques, our data has the form:

fracture	weight_kg	height_cm	waiting_time	bmd	No_Medication	Glucocorticoids	Anticonvulsant
NoFracture	64	155.5	18	0.8793	NA	NA	0.8793
NoFracture	78	162.0	56	0.7946	0.7946	NA	NA
NoFracture	73	170.5	10	0.9067	0.9067	NA	NA
NoFracture	60	148.0	14	0.7112	0.7112	NA	NA
NoFracture	55	161.0	20	0.7909	0.7909	NA	NA
NoFracture	65	168.0	7	0.7301	0.7301	NA	NA

And using the Amelia package in R to map out our missingness, the Missing Map below shows that there are a number of missing observations across the three variables, namely: No medication, Glucocorticoids and Anticonvulant. Since these missing observations couldn't be observed because of physical constrains, it implies our MDM is MAR. This dataset has 169 observations. We can evidently see that this map states that 25% of our data is missing. The rule of thumb in the Statistical community for missing data is that if 25% of it is missing, an appropriate number of imputation values would be 25 and above.



2 REFERENCE MODELS

In this section I run three dummy models, each for three different treatment levels which will serve as reference to assess if whether our standard errors will increase or decrease. If

our imputation method is not terribly bad, then the standard errors of our models should decrease by practice.

The following table shows the model of explaining the group without medication using other variables. It shows the predictors of the model, estimates, the standard error, 95% Confidence Interval and the corresponding p value.

No Medication Linear Model

Predictors	Estimates	std. Error	CI	p
(Intercept)	0.04526	0.18371	-0.31816 – 0.40869	0.806
weight kg	0.00473	0.00090	0.00295 – 0.00651	**<0.001**
waiting time	-0.00136	0.00055	-0.00245 – -0.00028	**0.014**
[NoFracture]	0.17289	0.02103	0.13130 – 0.21449	**<0.001**
height cm	0.00215	0.00122	-0.00027 – 0.00457	0.081
Observations	136			
R^2 / R^2 adjusted	0.576 / 0.563			

Out of 169 observed cases, only 136 were used to model the relationship of not taking medication with other variables, the rest were deleted as they had missing observations. 57.6% of the variation in our data can be explained by this model. Weight is a significant factor where for every 1kg increase in weight, the bone mineral density(bmd) score of a participant who doesn't take medication would increase by 0.00473. An increase in waiting time(statistically significant) also results in -0.00136 decrease in the bmd score. Not having a fracture suggests that there is 0.17289 increase in the bmd score of participants, as it is statistically significant. Height is significant under 10% significance level.

Glucocorticoids Linear Model

Predictors	Estimates	std. Error	CI	p
(Intercept)	-0.86073	0.73590	-2.40099 – 0.67954	0.257
weight kg	0.00286	0.00288	-0.00317 – 0.00890	0.334
waiting time	0.00032	0.00315	-0.00627 – 0.00690	0.921
[NoFracture]	0.23248	0.09635	0.03082 – 0.43414	**0.026**
height cm	0.00794	0.00507	-0.00267 – 0.01856	0.134
Observations	24			
R^2 / R^2 adjusted	0.541 / 0.444			

The above model only accounts/explains 54.1% of the variation in the bmd scores of participants who were treated with Glucocorticoids. For every non-fracture a participant has, their estimated bmd score increases by 0.23248.

Anticonvulsant Linear Model

Predictors	Estimates	std. Error	CI	p
(Intercept)	-0.98828	0.66731	-2.84103 – 0.86448	0.213
weight kg	0.01726	0.00545	0.00212 – 0.03241	**0.034**
waiting time	-0.00013	0.00255	-0.00721 – 0.00694	0.960
[NoFracture]	0.07898	0.06284	-0.09550 – 0.25346	0.277
height cm	0.00387	0.00521	-0.01060 – 0.01835	0.499
Observations	9			
R^2 / R^2 adjusted	0.897 / 0.795			

Weight is the only statistically significant factor in this model. For every 1kg increase in weight, there is an estimated 0.01726 increase in an individual's bmd score. This model accounts for 89.7% variation in the bmd scores of the participants that took Anticonvulant. We have the model

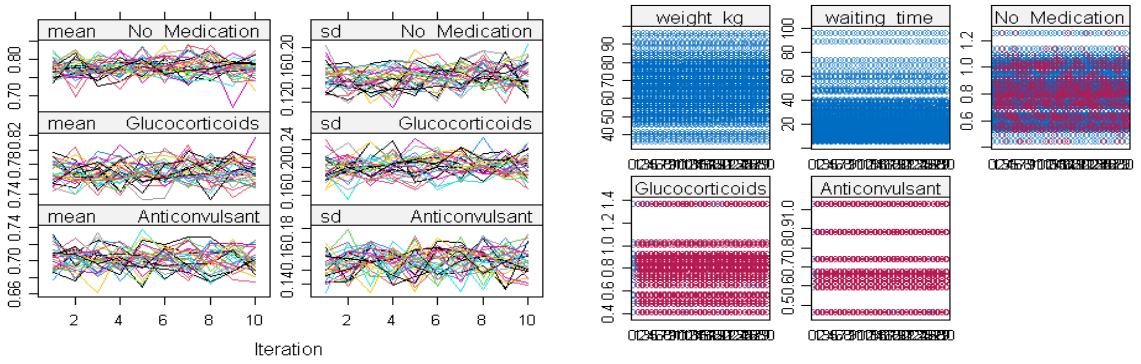
$$\text{Anticonvulant} = 0.00387\text{Height(cm)} + 0.07898\text{NoFracture} - 0.00013\text{WaitingTime} + 0.01726\text{Weight(kg)} - 0.98828$$

3 ANALYSIS

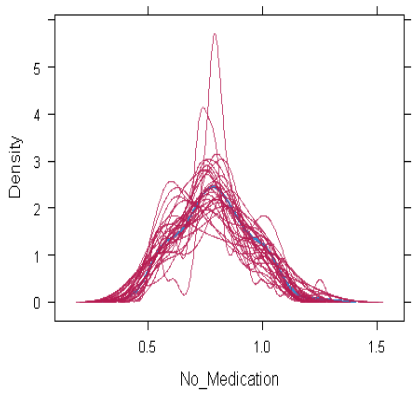
3.1 IMPUTATION

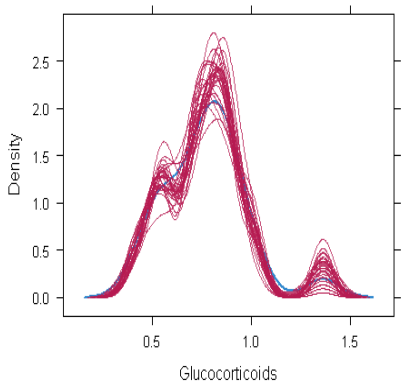
For this section, we will be using the mice package in R to run the MI algorithm. In this imputation algorithm, we first impute 30 data points with 10 iterations and a seed of 10, using the Predictive Mean Matching(PMM) method, sample and sample method for NoMedication, Glucocorticoids and Anticonvulant respectively. After the imputation, we observe how did the algorithm perform.

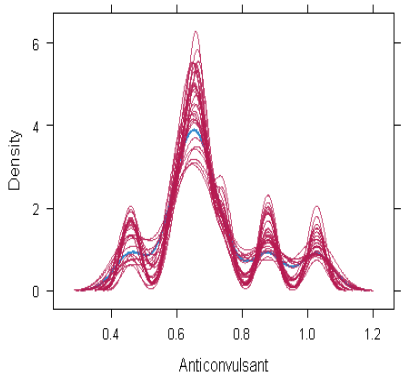
The plot below is the convergence plot(Bottom Left) of the three medication treatments that were imputed. This plot shows that the imputations of NoMedication are quick to converge, implying no issues with the imputations while the other two medication suggest otherwise for their cases. The imputed values align with the observed(Bottom Right).



The density plots below displays the individual imputations for the three different treatments with the Blue curve as the observed values & the Purple representing individual imputations. The imputations are close to the observed values with some variation while other imputed values are extreme.







3.2 IMPUTATION ASSESSMENT

The table below display the pooled models of the 30 imputations. These models only shows the fraction of missing information(fmi) and rate of imputation variance(riv). The Green values refer to estimates that are statistically significant at 5% while the Grey is significant at 10%. To look at the pooled models for the three treatments individually see Appendix A.

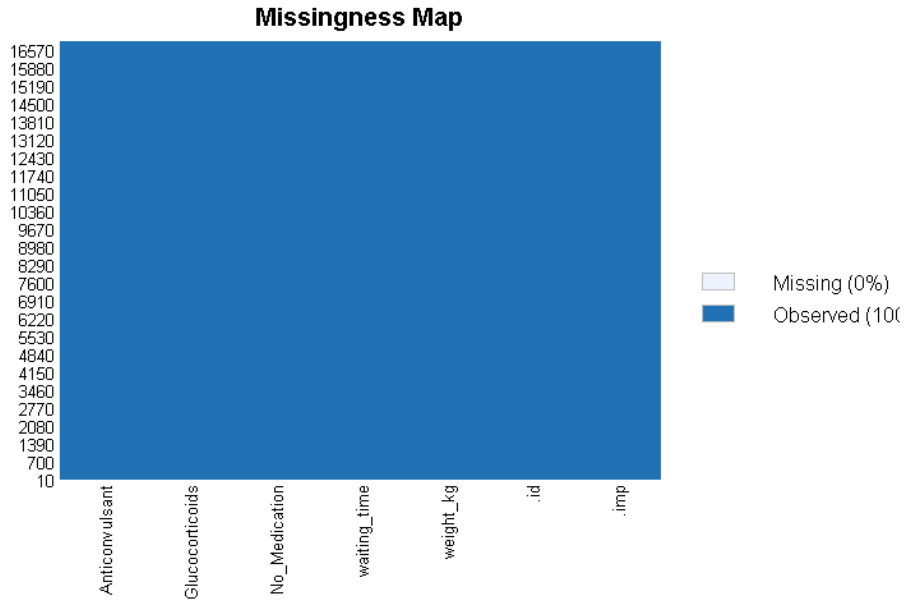
POOLED MODELS						
	No Medication		Glucocorticoids		Anticonvulsant	
term	**fmi**	**riv**	**fmi**	**riv**	**fmi**	**riv**
(Intercept)	0.23698	0.28602	0.50173	0.93144	0.53955	1.08155
weight_kg	0.23508	0.28294	0.58829	1.31420	0.61386	1.45895
waiting_time	0.16274	0.17565	0.52762	1.03169	0.53942	1.08099
NoFracture	0.20158	0.23097	0.48700	0.87875	0.58010	1.27150
height_cm	0.20565	0.23706	0.50942	0.96015	0.57191	1.23032

As seen from the above table, the fraction of missing information ranges from 16.274% – 23.698% for those without Medication, 48.7% – 58.829% for Glucocorticoids and 53.943% – 61.386% for Anticonvulsant. Imputations with fmi above 60% indicate that the

imputation method struggled and it is bad. RIV above 1.5 also indicate a struggling imputation with a lot of in-between variance.

3.3 IMPROVEMENT

This chapter is concerned with improving the previous imputation method by increasing the imputation number to 1000 and iterations to be 20. The observations are now 169 000. The following Missingness Map now shows that there is 100% of complete cases. The table that follows shows the new form of the dataset.



.imp	.id	weight_kg	waiting_time	No_Medication	Glucocorticoids	Anticonvulsant
1	1	64	18	0.7112	0.8829	0.8793
1	2	78	56	0.7946	0.5090	0.4586
1	3	73	10	0.9067	0.7215	0.6495
1	4	60	14	0.7112	0.8377	1.0287
1	5	55	20	0.7909	1.0020	0.6744
1	6	65	7	0.7301	0.9184	0.5899

The following table shows the corresponding pooled models for the three treatments after improvement. To look at the individual pooled models of the improved imputation see Appendix B. The imputations improved with lower fmi and riv values. The weight variable lost it's statistical significance while the intercept is slightly significant for Anticonvulant and No-Fracture loses its significant for Glucocorticoids. The standard errors have improved as they decreased.

POOLED MODELS						
	No Medication		Glucocorticoids		Anticonvulsant	
term	**fmi**	**riv**	**fmi**	**riv**	**fmi**	**riv**
(Intercept)	0.19003	0.21641	0.48674	0.90330	0.49627	0.93848
weight_kg	0.24211	0.29866	0.43972	0.74692	0.49851	0.94694
waiting_time	0.13940	0.14585	0.49433	0.93120	0.51116	0.99610
NoFracture	0.18462	0.20846	0.46710	0.83468	0.50422	0.96882
height_cm	0.18973	0.21596	0.48165	0.88502	0.49380	0.92923

4 APPENDIX

APPENDIX A : FOR CHAPTER 3.2 IMPUTATION ASSESSMENT

sPooled Model for No Medication

term	estimate	std.error	p.value	fmi	riv
(Intercept)	0.07013	0.19973	0.72620	0.23698	0.28602
weight_kg	0.00509	0.00097	**0.00000**	0.23508	0.28294
waiting_time	-0.00131	0.00060	**0.03141**	0.16274	0.17565
fractureNoFracture	0.14687	0.02264	**0.00000**	0.20158	0.23097
height_cm	0.00194	0.00131	0.13952	0.20565	0.23706

Pooled Model for Glucocorticoids

term	estimate	std.error	p.value	fmi	riv
(Intercept)	0.43246	0.44232	0.33290	0.50173	0.93144
weight_kg	0.00056	0.00236	0.81317	0.58829	1.31420
waiting_time	-0.00044	0.00143	0.75749	0.52762	1.03169
fractureNoFracture	0.02972	0.05055	0.55911	0.48700	0.87875
height_cm	0.00181	0.00297	0.54604	0.50942	0.96015

Pooled Model for Anticonvulsant

term	estimate	std.error	p.value	fmi	riv
(Intercept)	0.66311	0.35525		0.53955	1.08155
weight_kg	0.00022	0.00189	0.90914	0.61386	1.45895
waiting_time	0.00031	0.00112	0.78500	0.53942	1.08099
fractureNoFracture	0.00778	0.04301	0.85727	0.58010	1.27150
height_cm	0.00008	0.00245	0.97535	0.57191	1.23032

APPENDIX B : FOR CHAPTER 3.3 IMPROVEMENT

Pooled Model for No Medication

term	estimate	std.error	p.value	fmi	riv
(Intercept)	0.09789	0.19469	0.61592	0.19003	0.21641
weight_kg	0.00527	0.00098	**0.00000**	0.24211	0.29866
waiting_time	-0.00127	0.00059	**0.03387**	0.13940	0.14585
fractureNoFracture	0.14261	0.02249	**0.00000**	0.18462	0.20846
height_cm	0.00171	0.00130	0.18945	0.18973	0.21596

Pooled Model for Glucocorticoids

term	estimate	std.error	p.value	fmi	riv
(Intercept)	0.47705	0.43859	0.27986	0.48674	0.90330
weight_kg	0.00041	0.00205	0.84092	0.43972	0.74692
waiting_time	-0.00032	0.00139	0.81682	0.49433	0.93120
fractureNoFracture	0.02943	0.04990	0.55692	0.46710	0.83468
height_cm	0.00157	0.00291	0.59120	0.48165	0.88502

Pooled Model for Anticonvulsant

term	estimate	std.error	p.value	fmi	riv
(Intercept)	0.59951	0.34326	***0.08446***	0.49627	0.93848
weight_kg	0.00019	0.00168	0.90911	0.49851	0.94694
waiting_time	0.00012	0.00109	0.91564	0.51116	0.99610
fractureNoFracture	0.00552	0.04009	0.89083	0.50422	0.96882
height_cm	0.00052	0.00228	0.81932	0.49380	0.92923