BS845 Final Project – Jessy Yang

Multiple Imputation on Missing Values

Introduction

Addressing missing values is a critical aspect of any data analysis, particularly in epidemiological and clinical studies where data completeness directly impacts the reliability and validity of findings. Common measures to deal with missing values includes listwise-deletion and mean/median substitution, where the missing data is removed from the analysis or is replaced by the mean/median of the existing data points. These strategies may seem like an easy and quick fix, but at the same time carry potential issues in the analysis. Deleting datapoints reduces the degrees of freedom and statistical power, and replacing missing values with a fixed value decreases variance and could cause bias in the analysis.

The utilization of multiple imputation techniques holds significant value because it provides a systematic and statistically sound approach to handle missing data by imputing plausible values multiple times, thus capturing the uncertainty associated with the missing information. This methodology not only preserves the sample size but also produces more accurate and unbiased parameter estimates by maximizing the utility of available information.

This project will implement multiple imputation on a study dataset with missing values, assuming missing at random (MAR), and provide details on the imputation modelling used, such as number of imputed datasets that were created, variables included in the imputation procedure, and tests on normality assumption included in the model. Results (parameter estimates) for dataset with complete data, complete case analysis, single mean imputation, and multiple imputation will be compared to assess the bias improvement by addressing missing data. This project aims to demonstrate multiple imputation on missing data, in the pursuit of comprehensive and rigorous data analysis, leading to more robust and reliable conclusions in the face of incomplete data.

Methods

Framingham Heart Study Dataset was used for this project, with an original research topic of examining how risk factors predict cardiovascular disease. Data simulation and statistical analyses were carried out using R, with four sequential stages involved as listed below:

1. Data generation: A subset of the dataset including all the cardiovascular diseases (CVD) risk factors under consideration was created; missingness was checked to ensure a complete dataset was obtained. A logistic regression model was fitted, collinearity and goodness of fit were tested, and the effect estimates observed from the model were considered our reference results (with no missing data).
2. Amputation: the complete dataset was made incomplete, with 10%, 20%, and 40% missing observations of BMI variable, purposely only associated with variable sex to meet the Missing at Random (MAR) assumption for multiple imputation.[1]

   Since no guidelines has been established for an acceptable percentage of missing data which MI best benefits, the missingness of 10% to 40% was selected based on published

literatures suggesting that estimates are likely to be biased if more than 10% of data are missing but will only be considered hypothetic if the missing data are very large at 40%.[2-3]

3. Imputation: multiple imputation was performed for the simulated incomplete datasets; missing values were filled-in by two different imputation methods: the predictive mean matching (PMM) and random forest (RF) methods by the "mice" R package.

   All variables except for outcome CVD were included in the model as imputer for predicting missing values. The number of imputed datasets (M) was set to ten, which was suggested sufficient in most situations. Ten iterations were performed for each imputation within the MICE framework as recommended.[4] Also, for the RF-based imputation methods, the number of trees built was set to ten as suggested by previous study for less biased results.[5]

4. Analysis: statistical analyses were performed on both the original complete dataset and the imputed datasets, and comparisons of the different imputation methods were made.

   Logistic regression was performed on each imputed dataset by regressing CVD on the other variables in the dataset. The two imputation methods were compared for their accuracy of the imputed values using the normalized root mean squared error (NRMSE) and relative bias for the mean of the imputed variable. And the imputation methods were compared for their accuracy in estimating the effect estimates (Odds Ratio) in the logistic regression models.

Results

The complete dataset comprises data of age, sex, systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI), current cigarette smoking status, and diabetic status, and CVD event for 4415 subjects. 43.9% of the subjects were females. Average age was 50 years. Almost half (49%) of the subjects were current smokers, and 2.7% of the subjects were considered diabetic at examination. Subjects had average DBP and SBP of 83.1 and 132.9 and average BMI of 25.85.

Normality assumption was checked, and no independent variables had extreme skewness that needed transformation. Logistic regression model was fit, with all predictors showed statistically significant association with CVD event (all $p<0.05$). Multicollinearity was checked, and all independent variables had VIF <5. Test results are showed as below:

explanatory	OR(Complete Data)
AGE	1.06 (1.05-1.07, p<0.001)
SEX2	0.40 (0.34-0.46, p<0.001)
SYSBP	1.01 (1.01-1.02, p<0.001)
DIABP	1.01 (1.00-1.02, p=0.031)
CURSMOKE1	1.46 (1.25-1.71, p<0.001)
DIABETES1	3.24 (2.17-4.89, p<0.001)
BMI	1.03 (1.01-1.05, p=0.001)

Three incomplete data sets were then created with 10%, 20%, and 40% missing values for BMI. The missingness was make purposely to be dependent on variable sex but not on BMI itself to meet the missing at random (MAR) assumption, which is suggested for implementing multiple imputation. The MAR assumption was then back checked. (see Appendix Table 1 for all results)

Logistic regression model was fit for the three incomplete data sets, using the complete case analysis (CCA) that only included observations with complete data in the analyses. Notable biased OR estimates were observed for predictors DBP, smoking status, and diabetic status in all three CCA compared to model using complete dataset. (Table 1)

Table 1. Comparison of complete data and CCA with 10-40% MAR

label	levels	OR(Complete Data)	OR(CCA, na=10%)	OR(CCA, na=20%))	OR(CCA, na=40%))
AGE	Mean (SD)	1.06 (1.05-1.07, p<0.001)	1.06 (1.05-1.07, p<0.001)	1.05 (1.04-1.07, p<0.001)	1.07 (1.06-1.09, p<0.001)
SEX	1	-	-	-	-
	2	0.40 (0.34-0.46, p<0.001)	0.40 (0.34-0.47, p<0.001)	0.39 (0.32-0.46, p<0.001)	0.42 (0.34-0.51, p<0.001)
SYSBP	Mean (SD)	1.01 (1.01-1.02, p<0.001)	1.01 (1.01-1.02, p<0.001)	1.01 (1.01-1.02, p<0.001)	1.01 (1.01-1.02, p<0.001)
DIABP	Mean (SD)	1.01 (1.00-1.02, p=0.031)	1.01 (1.00-1.02, p=0.094)	1.00 (0.99-1.02, p=0.450)	1.01 (0.99-1.02, p=0.284)
CURSMOKE	0	-	-	-	-
	1	1.46 (1.25-1.71, p<0.001)	1.51 (1.28-1.78, p<0.001)	1.43 (1.20-1.70, p<0.001)	1.44 (1.18-1.77, p<0.001)
DIABETES	0	-	-	-	-
	1	3.24 (2.17-4.89, p<0.001)	3.50 (2.32-5.35, p<0.001)	3.85 (2.45-6.15, p<0.001)	4.43 (2.54-7.96, p<0.001)
BMI	Mean (SD)	1.03 (1.01-1.05, p=0.001)	1.03 (1.01-1.05, p=0.003)	1.04 (1.02-1.06, p=0.001)	1.04 (1.01-1.06, p=0.002)

Comparing the two imputation methods PMM and RF for 10%, 20%, and 40% missingness, PMM for dataset with 40% MAR had the smallest NRMSE and relative bias for the mean among the 6 models. PMM had a better accuracy in imputing missing values when the proportion of missingness is higher, whereas FR seemed to have better performance with lower proportion of missingness. No agreement on which method outperformed the other one regardless of the proportion of missingness was observed. (Table 2)

Table 2. Accuracy of the imputed values (in order)

Imputation model	NRMSE	Relative bias for the mean
PMM with 40% MAR	0.002940384	-0.0004666429
RF with 20% MAR	0.004243565	0.0006734595
PMM with 20% MAR	0.004426783	0.0007025365
RF with 10% MAR	0.004775163	0.0007578249
RF with 40% MAR	0.005620022	0.0008919051
PMM with 10% MAR	0.007393341	0.0011733333

Logistic regression effect estimates using data filled by PMM and RF methods were compared. In contrast to the accuracy of imputing missing values, the ability of the imputation methods to predict regression estimates suggested the opposite. For 10% MAR, dataset imputed by PMM generated an odds ratio of 1.46 for predictor smoking status closer to the “true value” from complete dataset (OR=1.46) compared to test results using dataset imputed by RF (OR=1.45). For 40% MAR, OR for predictor smoking status, diabetic status, and BMI from dataset imputed by PMM were all closer to the true values compared to RF. (see Appendix Table 2 for all results)

C-statistics were obtained for all seven logistic regression models: one complete dataset + six imputed. Area under the curve (AUC) for the fit model with the original non-missing data was 0.7418 (74.18%), and the second-best AUC of 0.7416 (74.16%) was observed in PMM with 10% MAR, followed by 0.7415 (74.15%) in RF with 10% MAR. There was a trend suggesting a better AUC when the imputed data was less (lower proportion of missingness) and a better AUC in PMM than in RF. However, the differences are very little. In terms of the accuracy of regression

model predictions it's hard to make conclusion on which model outperformed the other. We can only say that all models had a fair prediction power. (Table 3)

Table 3. Prediction power for the fitting logistic regression models (in order)

Imputation model	AUC (C-statistic)
Complete dataset	0.7418006
PMM with 10% MAR	0.7416066
RF with 10% MAR	0.7415421
PMM with 20% MAR	0.7413988
RF with 20% MAR	0.7413418
PMM with 40% MAR	0.7411216
RF with 40% MAR	0.7408117

When comparing the different imputation numbers (M=10, 20, and 50) in PMM, the difference in NRMSE and the relative bias for the mean were no more than 0.002, and the regression estimates were also similar across the imputed datasets. Therefore, it is considered that the increased imputation numbers had no significant effect on either the accuracy of the imputed values nor the accuracy of regression estimates. (see Appendix Table 3&4 for all results)

Regression estimates for non-missing data, complete-case analysis, single imputation using the mean, and PMM multiple imputation method were compared to confirm if multiple imputation was a better measure of addressing missing vales. For predictors "smoking status, diabetic status, and BMI", PMM had odds ratios that were the closest to the model using non-missing data. However, since the confidence intervals between these different methods were very similar and largely overlapped, we couldn't draw a conclusion that multiple imputation would provide significant improvement for this specific dataset. (Table 4)

Table 4. Comparison of CCA, SI, and MI

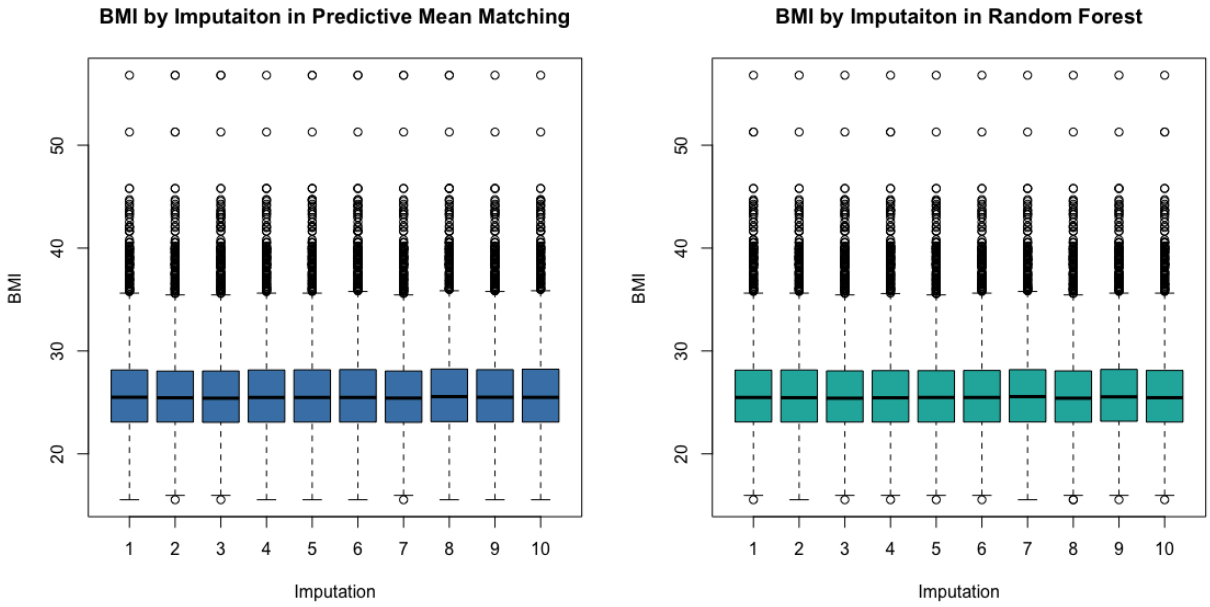
label	OR(Complete Data)	OR(CCA, na=40%))	OR(Single Imputation, mean)	OR (MAR 40%_PMM)
AGE	1.06 (1.05-1.07, p<0.001)	1.07 (1.06-1.09, p<0.001)	1.06 (1.05-1.07, p<0.001)	1.06 (1.05-1.07, p<0.001)
SEX				
	0.40 (0.34-0.46, p<0.001)	0.42 (0.34-0.51, p<0.001)	0.40 (0.34-0.46, p<0.001)	0.40 (0.34-0.47, p<0.001)
SYSBP	1.01 (1.01-1.02, p<0.001)	1.01 (1.01-1.02, p<0.001)	1.01 (1.01-1.02, p<0.001)	1.01 (1.01-1.02, p<0.001)
DIABP	1.01 (1.00-1.02, p=0.031)	1.01 (0.99-1.02, p=0.284)	1.01 (1.00-1.02, p=0.016)	1.01 (1.00-1.02, p=0.023)
CURSMOKE				
	1.46 (1.25-1.71, p<0.001)	1.44 (1.18-1.77, p<0.001)	1.44 (1.23-1.68, p<0.001)	1.45 (1.24-1.69, p<0.001)
DIABETES				
	3.24 (2.17-4.89, p<0.001)	4.43 (2.54-7.96, p<0.001)	3.26 (2.18-4.90, p<0.001)	3.22 (2.15-4.83, p<0.001)
BMI	1.03 (1.01-1.05, p=0.001)	1.04 (1.01-1.06, p=0.002)	1.04 (1.01-1.06, p=0.002)	1.03 (1.00-1.05, p=0.020)

Conclusion and Discussions

When comparing the two imputation methods PMM and RF in presence of MAR, a critical issue was that good predictive accuracy of imputed values did not necessarily mean better accuracy in estimating the associations. Random forest seemed to produce smaller NRMSE and bias when estimating the mean of the imputed values but predict more biased estimates compared to PMM. These results were likely to relate to how the two methods impute missing values, as PMM filled in the missing values with not predicted values but the observed data, whereas RF used the conditional mean for imputation by calculating the weighted average of the imputed

variable's observed values, resulting in underestimation of standard deviation and variance of the imputed variables. This finding was consistent with another research, suggesting that "imputation method that simply imputes missing values by minimizing prediction error can be problematic since it does not try to recover the joint distribution of the data and thus can result in biased parameter estimates." [6] However, from the box plots comparing the BMI distribution for the 10 imputed datasets between PMM and RF, RF did not show smaller variance. (Figure 1)

Figure 1. Distribution of BMI for 10 imputed datasets (PMM vs. RF; 20% MAR)



The paper also mentioned that the bias in regression estimates may be augmented based on different types of testing. It further concluded that "the performance of the imputation methods was weaker for logistic regression models because the binary outcomes provided less information for estimating regression parameters than the continuous outcomes, and the estimation of the log odds ratios was also more sensitive to inaccuracies in the imputed variables."

Another limitation is that the incomplete datasets were created only once for each missing rate. Since the model performance was tested on one dataset, it was not certain that we would get the same results on a different dataset with same proportion of missingness. Further steps are to generate a large number of datasets to be able to reduce the randomness and get a pooled result that is more representative from repeated testing.

In conclusion, this project did not find a clear advantage between PMM and RF in multiple imputation. One method may be more beneficial over the other based on different situations, such as the complexity of the variable types and the relationships among the covariates. Using multiple imputation improved biased estimates from complete-case analysis and single imputation, but the subtle difference did not provide statistically significant evidence for multiple imputation as the best measure of dealing with missing values.

Reference

1. Rubin, D. B. Inference and Missing Data. Biometrika. 1976;63(3):581–592.

2. Bennett DA. How can I deal with missing data in my study? *Aust N Z J Public Health*. 2001;25(5):464-469.

3. Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *J Clin Epidemiol*. 2003;56(1):28-37. doi:10.1016/s0895-4356(02)00539-5

4. Raghunathan TE, Solenberger PW, Van Hoewyk J., IVEware: Imputation and Variance Estimation Software, 2007 Ann Arbor, MI Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan (http://www.isr.umich.edu/src/smp/ive/).

5. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *Am J Epidemiol*. 2014;179(6):764-774. doi:10.1093/aje/kwt312

6. Hong, S., Lynn, H.S. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Med Res Methodol* **20**, 199 (2020). https://doi.org/10.1186/s12874-020-01080-1

Appendix I. Tables and Figures

Table 1. MAR assumption checks for three incomplete datasets (10%, 20%, 40% MAR)

Missing data analysis: BMI_mar0.1		Not missing	Missing	p
AGE	Mean (SD)	49.8 (8.7)	50.7 (8.4)	0.036
SEX	1	1784 (92.0)	155 (8.0)	<0.001
	2	2191 (88.5)	285 (11.5)	
SYSBP	Mean (SD)	132.7 (22.3)	134.9 (22.5)	0.046
DIABP	Mean (SD)	83.0 (12.1)	84.1 (12.0)	0.066
BMI	Mean (SD)	25.9 (4.1)	25.7 (4.0)	0.555
CURSMOKE	0	2003 (89.4)	238 (10.6)	0.155
	1	1972 (90.7)	202 (9.3)	
DIABETES	0	3862 (89.9)	434 (10.1)	0.096
	1	113 (95.0)	6 (5.0)	

Missing data analysis: BMI_mar0.2		Not missing	Missing	p
AGE	Mean (SD)	49.9 (8.7)	49.9 (8.7)	0.874
SEX	1	1629 (84.0)	310 (16.0)	<0.001
	2	1907 (77.0)	569 (23.0)	
SYSBP	Mean (SD)	132.8 (22.3)	133.3 (22.6)	0.526
DIABP	Mean (SD)	83.2 (12.0)	82.9 (12.1)	0.538
CURSMOKE	0	1771 (79.0)	470 (21.0)	0.079
	1	1765 (81.2)	409 (18.8)	
DIABETES	0	3439 (80.1)	857 (19.9)	0.781
	1	97 (81.5)	22 (18.5)	
BMI	Mean (SD)	25.9 (4.1)	25.7 (4.0)	0.166

Missing data analysis: BMI_mar0.4		Not missing	Missing	p
AGE	Mean (SD)	50.0 (8.8)	49.8 (8.5)	0.403
SEX	1	1319 (68.0)	620 (32.0)	<0.001
	2	1337 (54.0)	1139 (46.0)	
SYSBP	Mean (SD)	133.0 (22.3)	132.7 (22.4)	0.592
DIABP	Mean (SD)	83.3 (12.2)	82.8 (11.9)	0.239
CURSMOKE	0	1348 (60.2)	893 (39.8)	1.000
	1	1308 (60.2)	866 (39.8)	
DIABETES	0	2589 (60.3)	1707 (39.7)	0.438
	1	67 (56.3)	52 (43.7)	
BMI	Mean (SD)	25.9 (4.1)	25.7 (4.1)	0.114

Table 2. Comparison of PMM and RF

label	OR(Complete Data)	OR (MAR 10%_PMM)	OR (MAR 10%_RF)
AGE	1.06 (1.05-1.07, p<0.001)	1.06 (1.05-1.07, p<0.001)	1.06 (1.05-1.07, p<0.001)
SEX			
	0.40 (0.34-0.46, p<0.001)	0.40 (0.34-0.46, p<0.001)	0.40 (0.34-0.46, p<0.001)
SYSBP	1.01 (1.01-1.02, p<0.001)	1.01 (1.01-1.02, p<0.001)	1.01 (1.01-1.02, p<0.001)
DIABP	1.01 (1.00-1.02, p=0.031)	1.01 (1.00-1.02, p=0.026)	1.01 (1.00-1.02, p=0.024)
CURSMOKE			
	1.46 (1.25-1.71, p<0.001)	1.46 (1.24-1.70, p<0.001)	1.45 (1.24-1.70, p<0.001)
DIABETES			
	3.24 (2.17-4.89, p<0.001)	3.25 (2.17-4.88, p<0.001)	3.25 (2.17-4.87, p<0.001)
BMI	1.03 (1.01-1.05, p=0.001)	1.03 (1.01-1.05, p=0.005)	1.03 (1.01-1.05, p=0.008)

OR (MAR 20%_PMM)	OR (MAR 20%_RF)	OR (MAR 40%_PMM)	OR (MAR 40%_RF)
1.06 (1.05-1.07, p<0.001)	1.06 (1.05-1.07, p<0.001)	1.06 (1.05-1.07, p<0.001)	1.06 (1.05-1.07, p<0.001)
0.40 (0.34-0.46, p<0.001)	0.40 (0.34-0.46, p<0.001)	0.40 (0.34-0.47, p<0.001)	0.40 (0.34-0.46, p<0.001)
1.01 (1.01-1.02, p<0.001)	1.01 (1.01-1.02, p<0.001)	1.01 (1.01-1.02, p<0.001)	1.01 (1.01-1.02, p<0.001)
1.01 (1.00-1.02, p=0.028)	1.01 (1.00-1.02, p=0.028)	1.01 (1.00-1.02, p=0.023)	1.01 (1.00-1.02, p=0.017)
1.45 (1.24-1.70, p<0.001)	1.46 (1.24-1.70, p<0.001)	1.45 (1.24-1.69, p<0.001)	1.44 (1.23-1.69, p<0.001)
3.19 (2.12-4.78, p<0.001)	3.22 (2.15-4.82, p<0.001)	3.22 (2.15-4.83, p<0.001)	3.27 (2.18-4.89, p<0.001)
1.03 (1.01-1.05, p=0.004)	1.03 (1.01-1.05, p=0.005)	1.03 (1.00-1.05, p=0.020)	1.02 (1.00-1.05, p=0.021)

Table 3. Comparison of imputed values accuracy in PMM (M=10, 20, and 50; 40% MAR)

Imputation model	NRMSE	Relative bias for the mean
PMM (M=10)	0.002940384	-0.0004666429
PMM (M=20)	0.001492225	0.0002368181
PMM (M=50)	0.002272040	-0.0003605759

Table 4. Comparison of regression estimates accuracy in PMM (M=10, 20, and 50; 40% MAR)

label	OR(Complete Data)	OR (MAR 40%_PMM)	OR (MAR 40%_PMM_m=20)	OR (MAR 40%_PMM_m=50)
AGE	1.06 (1.05-1.07, p<0.001)	1.06 (1.05-1.07, p<0.001)	1.06 (1.05-1.07, p<0.001)	1.06 (1.05-1.07, p<0.001)
SEX				
	0.40 (0.34-0.46, p<0.001)	0.40 (0.34-0.47, p<0.001)	0.40 (0.34-0.46, p<0.001)	0.40 (0.34-0.47, p<0.001)
SYSBP	1.01 (1.01-1.02, p<0.001)	1.01 (1.01-1.02, p<0.001)	1.01 (1.01-1.02, p<0.001)	1.01 (1.01-1.02, p<0.001)
DIABP	1.01 (1.00-1.02, p=0.031)	1.01 (1.00-1.02, p=0.023)	1.01 (1.00-1.02, p=0.017)	1.01 (1.00-1.02, p=0.020)
CURSMOKE				
	1.46 (1.25-1.71, p<0.001)	1.45 (1.24-1.69, p<0.001)	1.44 (1.23-1.68, p<0.001)	1.44 (1.23-1.69, p<0.001)
DIABETES				
	3.24 (2.17-4.89, p<0.001)	3.22 (2.15-4.83, p<0.001)	3.25 (2.16-4.87, p<0.001)	3.24 (2.16-4.85, p<0.001)
BMI	1.03 (1.01-1.05, p=0.001)	1.03 (1.00-1.05, p=0.020)	1.02 (1.00-1.05, p=0.072)	1.02 (1.00-1.05, p=0.034)

Appendix II: R code (attached in a separate file)