Multiple imputation

Q Yun

April 2024

1. The data

In order to investigate the performance of the MICE package (van Buuren and Groothuis-Oudshoorn, 2011) and multiple imputation approach, the Boston Housing data in the MASS package (Venables and Ripley, 2002) will be used in this experiment. This dataset, with a total of 506 cases, contains information about the housing values in the suburbs of Boston MA area. There are 14 variables, which are crim, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, lstat and medv (detailed description of the variables is available in RStudio) in the dataset, and it doesn't contain any missing data.

Among the variables, *chas* is binary consisting of values of 1 and 0, and other variables are continuous with the exception of *zn* and *rad* being discrete. However, it is worth noting that many of them are ratio data, and none of the variables contains any negative values. An extract of the data is shown in table 1.

Table 1: Extract of the Boston housing dataset

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
0.01965	80	1.76	0	0.385	6.230	31.5	9.0892	1	241	18.2	341.60	12.93	20.1
0.10469	40	6.41	1	0.447	7.267	49.0	4.7872	4	254	17.6	389.25	6.05	33.2
0.06047	0	2.46	0	0.488	6.153	68.8	3.2797	3	193	17.8	387.11	13.15	29.6
0.01360	75	4.00	0	0.410	5.888	47.6	7.3197	3	469	21.1	396.90	14.80	18.9

2.The model

The main interest of our model is how the *medv* variable, which is the median house prices, can be predicted by other covariates, therefore *medv* is the dependent variable. While some covariates appear to be highly correlated in an initial analysis of the data, we include them all as the explanatory variables, for they might be useful in the multiple imputation of the missing data. However the possible interactions among them are not considered. The model can be expressed as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + ... + \beta_{12} x_{12} + \beta_{13} x_{13} + \epsilon$$

with y being the *medv* variable, x_1 to x_{13} being the explanatory variables of *crim*, *zn*, *indus*, *chas*, *nox*, *rm*, *age*, *dis*, *rad*, *tax*, *ptratio*, *black*, and *lstat* respectively. We fit this model and obtain the following estimates (Table 2) as the benchmark in assessing the performance of different multiple imputations.

Table 2: Estimates of coefficients in the benchmark model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.4594884	5.1034588	7.1440742	0.0000000
crim	-0.1080114	0.0328650	-3.2865169	0.0010868
zn	0.0464205	0.0137275	3.3815763	0.0007781
indus	0.0205586	0.0614957	0.3343100	0.7382881
chas	2.6867338	0.8615798	3.1183809	0.0019250
nox	-17.7666112	3.8197437	-4.6512574	0.0000042
rm	3.8098652	0.4179253	9.1161402	0.0000000
age	0.0006922	0.0132098	0.0524024	0.9582293
dis	-1.4755668	0.1994547	-7.3980036	0.0000000
rad	0.3060495	0.0663464	4.6128998	0.0000051
tax	-0.0123346	0.0037605	-3.2800091	0.0011116
ptratio	-0.9527472	0.1308268	-7.2825106	0.0000000
black	0.0093117	0.0026860	3.4667926	0.0005729
lstat	-0.5247584	0.0507153	-10.3471458	0.0000000

3. Exploratory data analysis

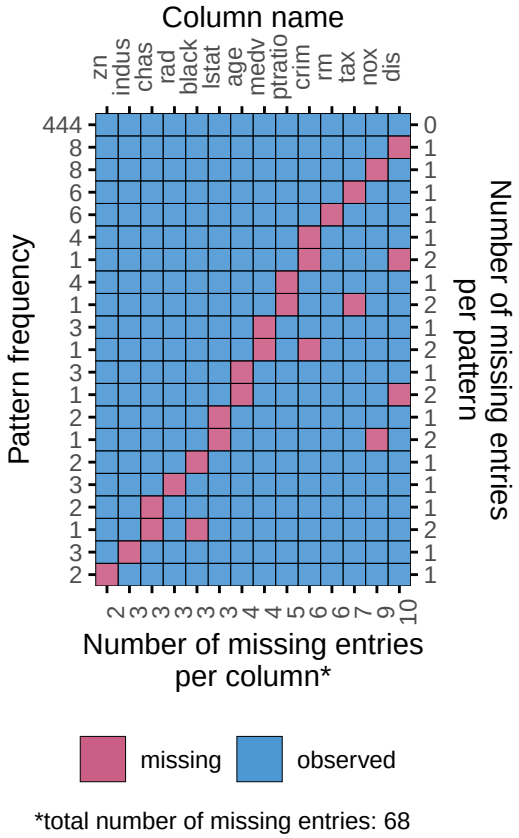

Figure 1: Missing pattern of the first dataset (missing 1% values)

We randomly assign 1% of the values in the dataset with NA, creating the first modified dataset for our exploratory analysis. For any dataset with missing data, it is essential to analyse the missing pattern to check if it is appropriate to implement multiple imputation. Although it is usually fine to implement it for the MCAR (missing completely at random) and MAR (missing at random) cases, it may be problematic for the NMAR (not missing at random) cases.

As can be seen in Figure 1, in this modified dataset with 1% missing data, 444 observations are fully complete, representing 87.7% of the data. The largest proportion, 8 records representing about 1.6% of the data, misses only the values for the variable of *dis*, and there are another 8 records with the value of *nox* missing. There are only 6 records with two variables missing, so it seems that the missingness of one variable is not linked to the missingness of any other variable. The majority of records with missing data, 10 records representing about 2% of the data, do not have information on the variable of *dis*.

If we had no idea about the missingness of this dataset, we may want to explore whether the missingness of *dis* (weighted distances to five Boston employment centres), is related to *rad* (index of accessibility to radial highway) by plotting a histogram of *rad*, split by whether *dis* is missing or not.

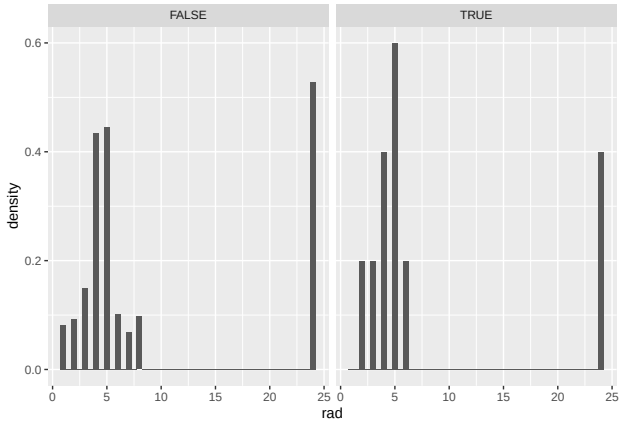

Figure 2: Exploring the relationship of missingness of rad to dis

From the above histograms (Figure 2), the distributions of *rad* appear to similar whether *dis* is missing or not, which indicates that the missingness of *dis* is not linked to *rad*.

4.Method

We introduce three different proportions of missingness (1%, 10%, 20%) to the data, creating three datasets for our experiments. For each dataset, two different built-in imputation methods in the mice package (Burren and Groothuis-Oudshoorn, 2011) are implemented for the comparison of their coefficients estimates. Certain built-in method in MICE, such as norm, might not be appropriate for this dataset when we consider that some variables are binary and some others, such as *indus*, *age* and *lstat*, are percentages. In addition, none of the variables contain any negative values. For the number of imputations, as higher percentage of missingness tends to require more iterations, we generally follow the recommendation that the amount of iterations should be at least equal to the percentage of missing observations (White et al.,2011). The experiment will be carried out in the following steps:

1. Generate three datasets, each with a different proportion of missingness (1%, 10% and 20% respectively) by randomly assigning certain values to NA;
2. Carry out a complete case analysis on each dataset;
3. For each dataset, use MICE package to carry out the multiple imputation with two methods, the predictive mean matching (PMM) and the classification and regression trees (CART) with a certain number of iterations. Check the convergence of the algorithm and the feasibility of imputed data before the pooled analysis.
4. Treat the coefficients from each pooled analysis as a vector, and then compare them to the results of the complete case analysis against the benchmark by using the idea of Euclidean distance. The estimates with the shorter distance is a better estimate.

5.Convergence of the MICE algorithm

For the first dataset with 1% missing data, we use MICE to generate 10 datasets with 5 iterations, by firstly using the PMM method, and investigate the convergence of the algorithm.

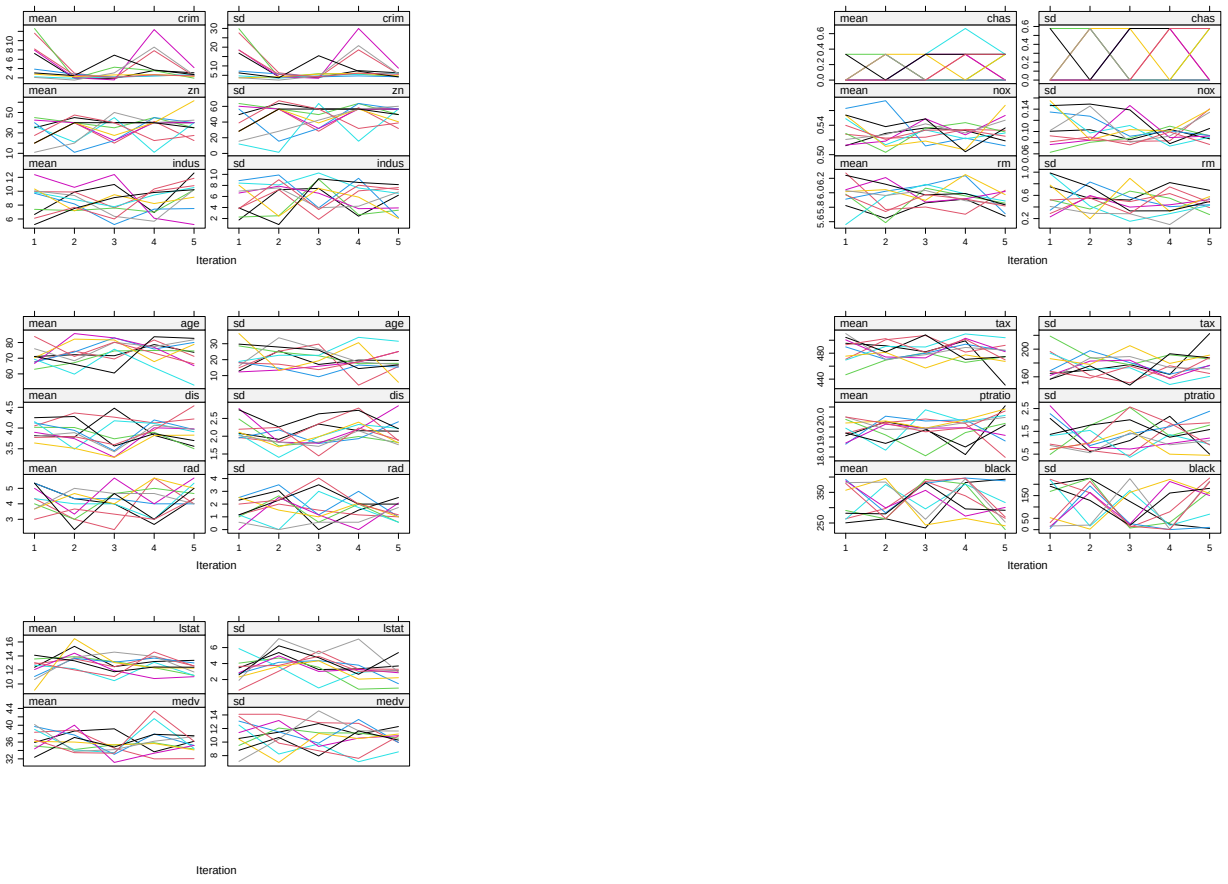

Figure 3: The mean and standard deviation of missing variables over the internal iterations of the MICE algorithm

By plotting the imputed data (see Figure 3), 5 iterations already converge nicely for this small percentage of missingness (further analysis for the CART method also shows convergence). We

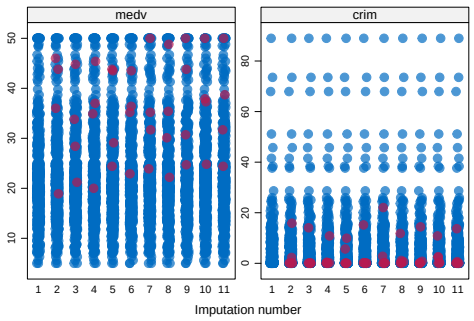

Figure 4: Strip plot of imputed variables of medv and crim (the observed data is shown in blue and imputed data in red)

then examine if the imputed data is plausible by looking at the summary of a couple of imputed datasets, particularly the minimum and maximum values, for an initial analysis, and they appear to be satisfactory. Another examination by using stripplot (by checking the imputed value for *mediv* and *crim*) also seems to support that the imputed data are plausible (see Figure 4).

6. Results

Since the imputed data are plausible and the algorithm converges nicely, we use the imputed data for our model and pool the analysis together, then compare them to the benchmark by calculating the Euclidean distance, with the same analysis being applied to the results of the complete case analysis.

By following the same steps, we generate two more datasets with 10% and 20% respectively for our analysis. For these two percentages of missingness, we increase the number of iteration in the MICE imputation to 30 and 45 times respectively. After checking the convergence of the algorithms and the distribution of imputed data, we decide to apply the imputed data to our model for pooled analysis as well.

Table 3: Euclidean distance of different imputations to the benchmark for data with different percentage of missingness

missingness_percent	Euclidean_complete_case	Euclidean_pmm	Euclidean_cart
5%	0.4960091	1.058774	0.2144800
10%	24.1248839	1.932664	0.9346286
20%	51.0088303	2.006520	0.2806388

It can be seen from Table 3, when the percentage of missingness is low, there isn't much difference between the Euclidean distances of the three analyses, and the complete case analysis is satisfactory. However, when this percentage increases to 10% or 20%, complete case analysis performed rather poorly, while the performance of the two imputation methods is far more stable, showing their superiority over complete case analysis. Between the PMM and CART methods, the Euclidean distance of CART method seems to perform a bit better than the PMM method for our data.

In order to check if the results of pooled analysis are satisfactory, we also need to check the *lambda* (proportion of variance due to the missing data) and *fmi* (fraction of missing information) to see the impact of missing data and mutliple imputations on the variance of the estimates. The maximum values of *lambda* and *fmi* in Table 4 suggest that most of them are under 0.08, although the values for the PMM method fall between 0.12 to 0.16 when the percentage of missingness increases to 10% and 20%, which also seems to suggest that the CART method works better for our data.

Table 4: Maximum value of lambda and fmi in different imputations

missingness_percent	pmm_lambda	pmm_fmi	cart_lambda	cart_fmi
5%	0.0726461	0.0777724	0.0436739	0.0481295
10%	0.1490753	0.1572400	0.0524402	0.0570667
20%	0.1245193	0.1315321	0.0279029	0.0321250

7. Conclusion

From our experiment, it can be seen that multiple imputation in MICE performs much better than the complete case analysis, particularly when the percentage of missingness increases. The variance of coefficient estimates due to the missing data and imputation remains fairly stable during the pooled imputation analysis. Therefore, multiple imputation is recommended for the analysis of missing data of MAR or MCAR cases, although we should bear in mind that the amount of iteration needs to increase when the percentage of missingness increases and different imputation methods may performs differently.

8.References

Allaire J, Xie Y and Dervieux C, et al. (2023). *rmarkdown: Dynamic Documents for R*. R package version 2.25, https://github.com/rstudio/rmarkdown.

Oberman H (2023). *ggmice: Visualizations for 'mice' with 'ggplot2'*. R package version 0.1.0, https://CRAN.R-project.org/package=ggmice.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

van Buuren,Stef and Groothuis-Oudshoorn, Karin (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. DOI 10.18637/jss.v045.i03.

Wickham H, Averick M, and Bryan J et al (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686 https://doi.org/10.21105/joss.01686.

Wickham H, François R, Henry L, Müller K, Vaughan D (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4, https://CRAN.R-project.org/package=dplyr.

White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377–399. doi: 10.1002/sim.4067.