Multiple Data Imputation - MICE

Noa Miller

2024-04-01

Introduction

The objective of this project is to experiment with multiple data imputation (MICE) and compare the results of a linear model fit to the data. The data set contains medical insurance cost of 2772 observations of patients aged between 18 and 64 years. It contains various factors influencing medical insurance charges and aims to enhance forecasting medical expenses for policyholders. Categorical variables are sex (51% male and 49% female), smoker status (20% smoker) and the region (southwest, southeast, northwest, northeast - nearly evenly stratified). Continuous variables are charges and bmi (Body Mass Index), discrete variables are age and the number of children. The variable of interest is medical insurance charges. The original data set is complete and contains no missing values. In order to experiment with the MICE algorithm, the data will be intentionally modified to contain missing values by randomly deleting observations across each variable.

The Model

A linear regression with three factor variables (sex, smoker and region) is fit to the data:

$$Y_{i,j,k,l} = \mu + \gamma_j + \theta_k + \tau_l + \alpha * x_{i,1} + \beta * x_{i,2} + \delta * x_{i,3} + \epsilon_{i,j,k,l} \text{ where}$$

$Y_{i,j,k,l}$ is the log of the insurance charge of observation i for i = 1, 2, ..., 2772 within gender group j, smoker group k, region group l.

μ is the expected overall mean for a non-smoker female patient from the region northeast.

γ_j: mean effect of sex $\gamma_j = 0$ for female j = 0 and $\gamma_1 = 1$ for male patients, j = 0, 1.

θ_k: mean effect of smoke $\theta_k = 0$ for non-smokers k = 0 and $\theta_1 = 1$ for smoker patients, k = 0, 1.

τ_l: mean effect of region $\tau_l = 0$ for patients from the region northeast, otherwise τ_l for l = 1, 2, 3 where τ_1 represents northwest, τ_2 southeast and τ_3 southwest respectively.

α: mean effect of age in years.

β: mean effect of bmi.

δ: mean effect of number of children.

with the assumption that the error term $\epsilon_{i,j,k,l} \sim \text{Normal}(0, \sigma^2)$ normally distributed with common variance σ^2 of fitted values.

Exploratory Data Analysis (EDA)

There is a relatively strong covariance between charges and bmi, as well as charges and age.

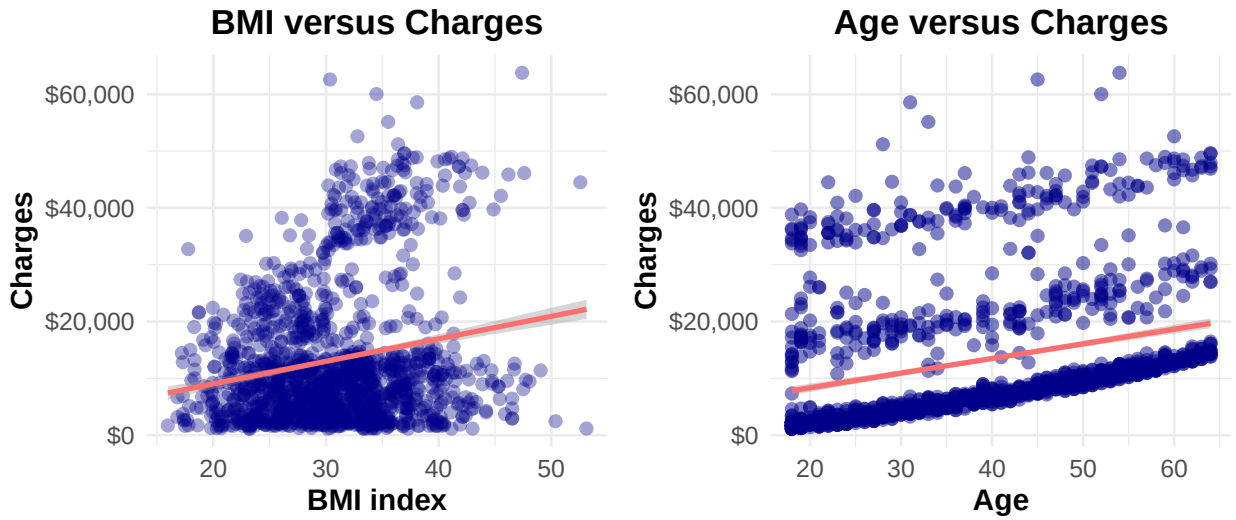

Figure 1: Charges tend to linearly increase as BMI index surges. This trend is observed in terms of age as well. However, three distinct price categories can be determined which are determined by other factors than age, such as smoker status or the number of children.

Further to that the visual inspection of the data suggests that smoker status will play an important role in determining insurance charges. On the other hand, the exploratory analysis showed that patients with no children tend to pay higher medical insurance cost, and sex or region play a rather insignificant role. Therefore from a modeling perspective it might be sensible to remove sex and region from the model, but we will keep these variables for now and focus on imputation. To prevent μ being negative value, the log of insurance charges was fitted in the regression model.

To carry out statistical analysis with missing data, the missingness pattern was analyzed, since MICE assumes and performs best when data is missing at random. Missing values were introduced by hand and completely at random. Therefore the EDA focuses on visualizing the pattern of missing data points across variables, rather than the mechanism of missingness (MCAR, MAR, NMAR).

Table 1: Missingness probability across variables tend to defined benchmark of 5%.

age	sex	bmi	children	smoker	region	charges
0.05	0.043	0.054	0.043	0.046	0.049	0.057

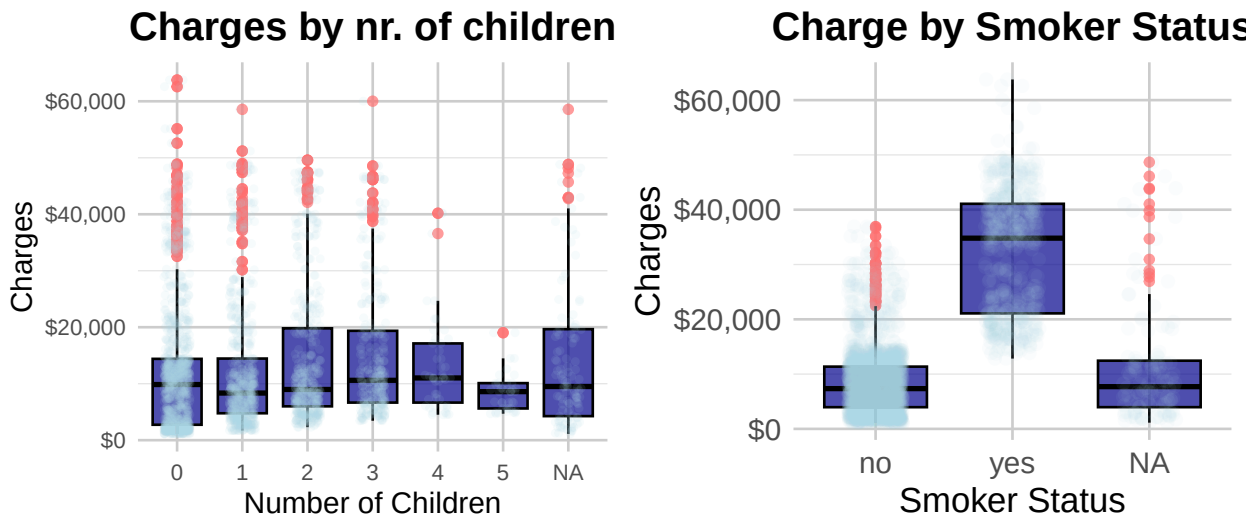

Figure 2: Jittered points represent individual observations. Most patients have no children, and their rate tend to be the highest. However, largest median charge is displayed by patients having three children. Smoking is definitely significant for determining insurance costs. NA values got well spread by and are representativ of the distribution in terms of smoker status.

Missing values got distributed relatively well spread across charges and other continuous variables, as well as within smokers and non-smokers. Exploring these distributions is reassuring since some explanatory variables (smoker, bmi, age) strongly correlate with the dependent variable. Further visual analysis confirmed no major patterns in the data and it can be concluded that missingness of observations is completely at random (MCAR), as expected.

Method

Imputation was carried out with the MICE algorithm at m = 5 imputed sets first, and in case of unsuitable convergence, the number of imputed sets got increased to 10 in total. For factor variables the method “polyreg” (Polynomial Regression) is being set. For continuous variables the method “norm” (Gaussian imputation) and “rf” (imputation with Random Forest model) was tested.

Three test sets got defined with the following proportion of missing data:

- proportion of missingness at 5% (benchmark set)
- proportion of missingness at 20%
- proportion of missigness at 40%

Imputed values will be analyzed on each of the three test sets to compare the performance of the two methods (normal vs rf). The quality of imputation was monitored by visualizing the imputation chains for each of the three sets. Finally, the main focus of the analysis was given to fitting the regression model and comparing parameter estimates and standard errors between:

- the full data set
- the modified data set (at 5% missingness) versus complete data (NA’s omitted)
- the modified data set versus multiple imputation with “random forest”rf” and “polyreg” method on the benchmark set (5% missingness).

Convergence of the MICE algorithm

As diagnostic plots trace lines of the MICE algorithm were produced for convergence evaluation on each of the three test sets (5%, 20%, 40%).

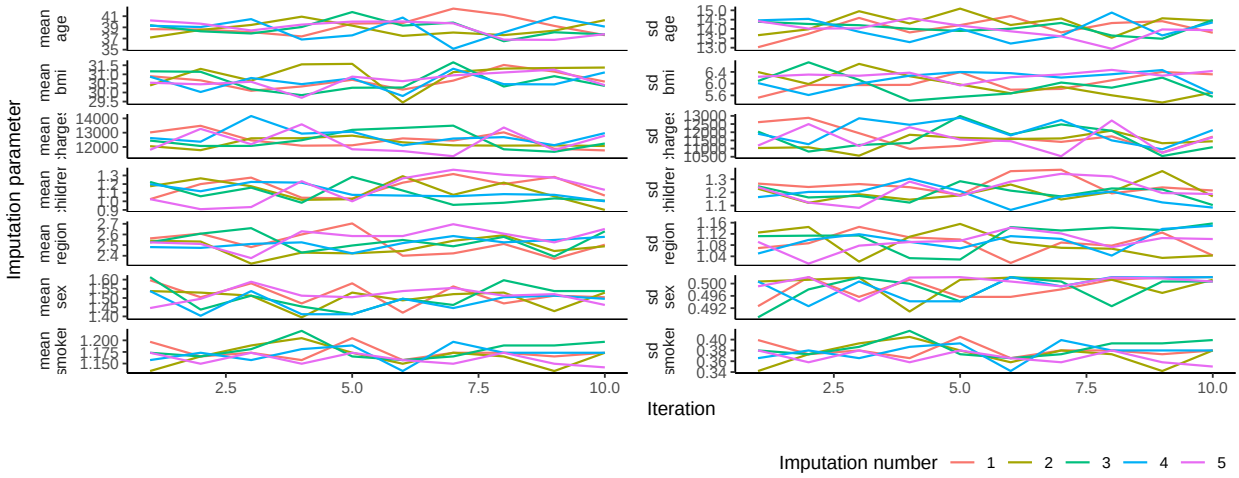

Figure 3: MICE trace plots of the benchmark set (5%) imputed with random forest and polyreg combination. Each colored chain represent an imputed set and there is no particular pattern in the traces. They converge properly and 10 iterations seem sensible.

In terms of the number of iterations 10 proved to be sensible in each case of the test sets. The trace plot lines represent imputed values across the iterations. Each chain is a single iteration and only the final value gets taken over to the imputed data set.

Results

The analysis has shown that the combined "normal" & "polyreg" versus "rf" & "polyreg"method of MICE performs similar on the data in terms of model output.

Table 2: Param. est. with imputation normal vs random forest.

	μ	α	γ_1	β	δ	θ_1	τ_1	τ_2	τ_3
norm	7.066	0.033	-0.087	0.015	0.098	1.557	-0.053	-0.173	-0.124
rf	7.149	0.032	-0.077	0.013	0.102	1.521	-0.059	-0.172	-0.127

Table 3: St. errors with imputation normal vs random forest.

	μ	α	γ_1	β	δ	θ_1	τ_1	τ_2	τ_3
norm	0.056	0.001	0.022	0.002	0.008	0.025	0.030	0.030	0.032
rf	0.067	0.001	0.021	0.002	0.009	0.034	0.029	0.032	0.028

However, it is still important to consider unrealistic imputations the MICE algorithm with normal method produced on missing data. The analysis showed that the normal imputation method does not differentiate between discrete and continuous values, and by taking the Gaussian mean and standard deviation of observed values unrealistic data (such as 2.25 number of children) can be generated. As a result fitting a model to such imputed values can lead to distorted parameter estimates and likely artificially lower standard error values.

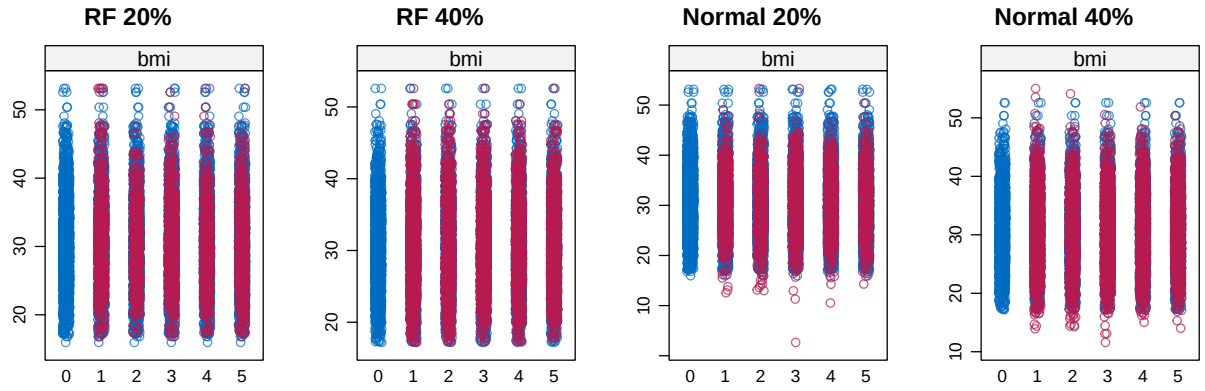

Figure 4: BMI values with the two imputation methods. The blue dots are observed values, red dots the imputed ones.

As a consequence, it is more sensible to chose the "random forest" method, as it fits the model to realistic imputed variables.

In terms of the main analysis there is no difference in parameter estimates between the modified benchmark data (5% probability of missingess) and the completed set (na.omit()). The parameters of the model fitted to the imputed data are somewhat different to the model fitted to the full set, but the values are relatively close to each other.

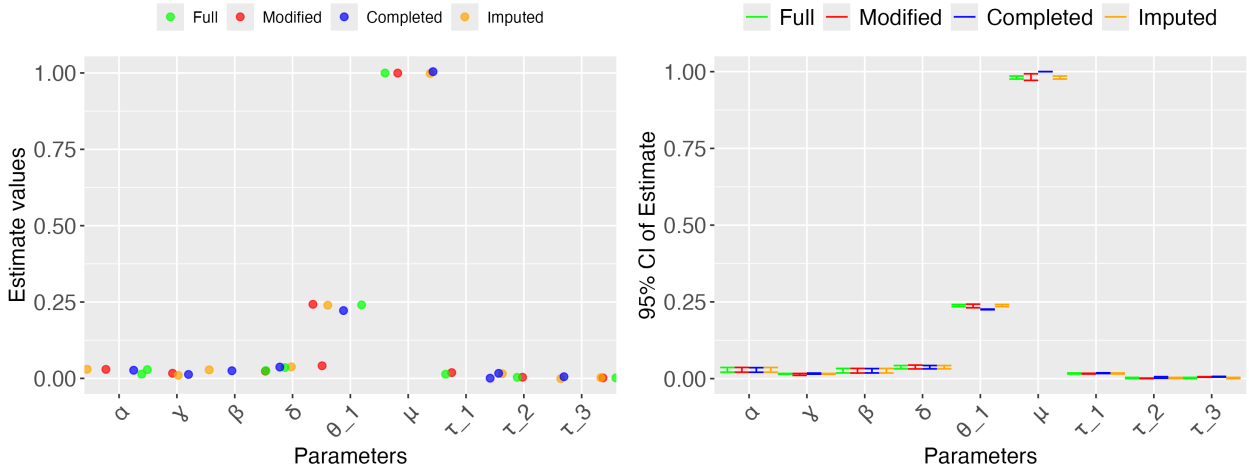

Figure 5: Scaled parameter estimates, and their respective 95% Confidence Intervals.

The 95% confidence intervals were calculated with the quantile function of a t-distribution with 2771 degrees of freedom qt(0.975, 2771), which could be very well approximated by a normal distribution since n is large enough. All intervals of parameter estimates overlap, however the interval of the intercept and smoker variable are somewhat different for the full and imputed sets. This might be a borderline significance, the imputed set might underestimate the importance of smoker and slightly overestimate the intercept. However overall it can be concluded that the difference is estimate for the rest of the parameters is not significant. The model fitted to the full data provides least uncertainty about the parameter estimates, followed by the imputed model, where standard errors are somewhat higher. Regarding the modified and completed sets standard errors are identical, just like for parameter estimates.

Table 4: Comparison of st. errors between the four sets

	μ	α	γ_1	β	δ	θ_1	τ_1	τ_2	τ_3
full	0.051	0.001	0.017	0.001	0.007	0.021	0.025	0.025	0.025
modified	0.061	0.001	0.021	0.002	0.008	0.026	0.030	0.030	0.030
completed	0.061	0.001	0.021	0.002	0.008	0.026	0.030	0.030	0.030
imp. rf	0.067	0.001	0.021	0.002	0.009	0.034	0.029	0.032	0.028

Conclusion

The MICE algorithm is a powerful tool to help conduct statistical analysis on missing data as it takes into account the relationship between variables when imputing values. This analysis explored that choosing the right imputation method is key to ensure a statistical model is fitted to realistic data values and generate sensible prediction estimates for accurate forecasts. If the chosen MICE imputation method is unfit to the data at hand, imputed values might be unrealistic, regardless of the proportion of missing values in the original set. This can lead to less reliable prediction accuracy of a model.

There will likely be little or no difference is terms of model output when the proportion of missing data is low (e.g. 5%), if MICE imputation is done properly. Diagnostic plots are useful tools to track the quality of imputed values. For data with significant proportion of missingness (40%) MICE is still capable of producing high quality imputations. When properly used, it is a reliable tool generating realistic, observed-like data for any set that has missing observations.

References

https://www.kaggle.com/datasets/rahulvyasm/medical-insurance-cost-prediction/.

https://cran.r-project.org/web/packages/ggmice/vignettes/ggmice.html/.

https://search.r-project.org/CRAN/refmans/mice/html/mice.html/.

https://cran.r-project.org/web/packages/mice/mice.pdf.

https://cran.r-project.org/web/packages/miceRanger/vignettes/miceAlgorithm.html.

https://link.springer.com/article/10.1007/s11121-007-0070-9.

https://bookdown.org/yihui/rmarkdown-cookbook/kable.html.

https://www.rdocumentation.org/packages/kableExtra/versions/1.4.0/topics/column_spec.

https://www.compart.com/en/unicode/search?q=alpha#characters.

University of Sheffield, MAS61006 Bayesian Statistics and Computational Methods Lecture Notes.

(Van Buuren and Oudshoorn 2000)
Van Buuren, Stef, and Catharina GM Oudshoorn. 2000. "Multivariate Imputation by Chained Equations." Leiden: TNO.