MAS61006 Assessed Project

220209225

April 2024

The data

The data used is a simulated data of medical expenses for patients in the USA using demographic statistics from the US Census Bureau with 1338 observations, consisting of **age** (Integer, age of primary significance excluding above 64 year olds), **sex** (Binary character, male, female), **bmi** (Continuous number, Body Mass Index (kg / m^2) of policy holder), **children** (Integer 0 - 5, number of children/dependents covered by insurance plan), **smoker** (Binary character (yes, no), regular smoker or otherwise), **region** (4 level factor of northeast, northwest, southeast, southwest), beneficiary place of residence) and **charges** (Continuous number, the cost of insurance to the beneficiary[1])

The model

Linear regression assumptions were verified (see code), revealing that applying a log transform to the charges response variable aligned with the requirements of linear regression. The other variables were treated as explanatory variables. The model is noted below:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \beta_7 x_{3i} x_{5i} + \epsilon_i$$

Where:

- y_i - the log transformation of the charges variable
- x_{1i} - age
- x_{2i} - sex ($x_2 = 1$ if the sex is male)
- x_{3i} - BMI
- x_{4i} - children ($x_4 = 1$ if the beneficiary has children)
- x_{5i} - smoker ($x_5 = 1$ if the beneficiary is a regular smoker)
- x_{6i} - region ($x_6 = 1$ if the beneficiary lives in the northwest, southeast or southwest)
- $\beta_7 x_{3i} x_{5i}$ - the interaction term between BMI and smoker status (see EDA section)

EDA

The dataset initially had no missing data, but it was manipulated using a script to artificially introduce a 15% missingness randomly, a frequency often observed in surveys/censuses[2]. Further exploration of the 'children' and 'region' variables showed no significant patterns, and the distribution of charges across regions was strikingly similar, undermining any potential correlation with the number of children. An interaction effect between BMI and smoker status further necessitates the exploration of various imputation methods, as some might better capture this complex relationship.

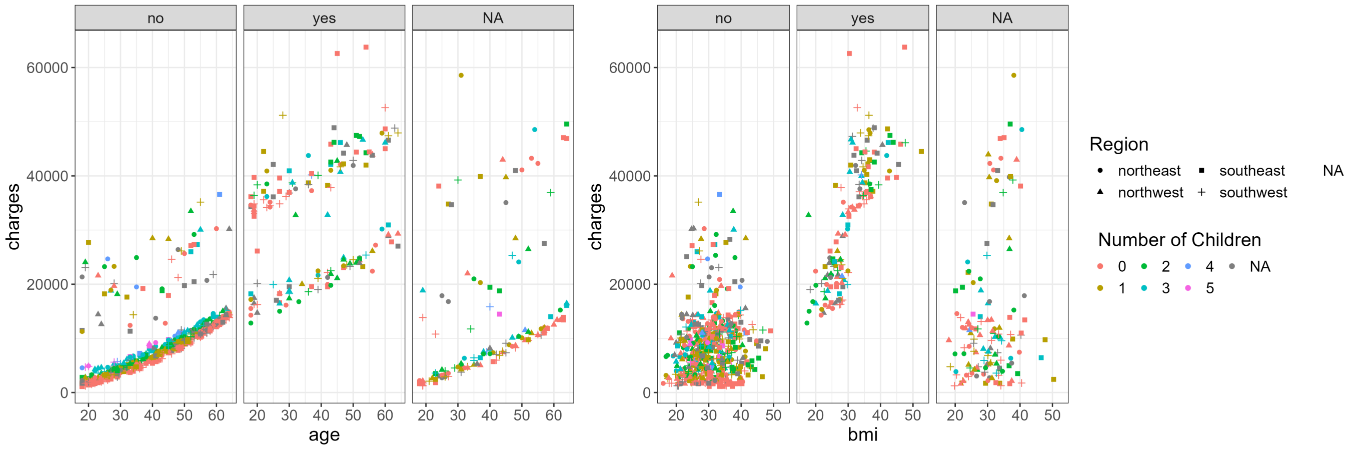

Figure 1: Age, BMI, and Insurance Premiums across Smoking Statuses (Yes, No, NA). Non-smokers show a notable positive correlation between age and premiums, while smokers tend to have higher premiums with positive associations with BMI and age. An interaction effect emerges for BMI among smokers, indicating a positive correlation with premiums, contrasting with non-smokers. The data reveals no correlation between variables and missing values.

Therefore, we will focus on analysing the MICE algorithm's performance under varying conditions: different levels of missingness (**Experiment A**), alternative imputation methods (**Experiment B**), and robustness to varied missingness sampling (**Experiment C**).

Method

Unless specified, all plots will use a seed of 42 with 15% missing data, along with the PMM default imputation method.

Experiment A

1. Using `set.seed(42)`, generate 3 datasets with varied probabilities of missingness (0.05, 0.15, 0.25).
2. For each dataset, fit the linear model specified in the model section to perform a complete case analysis. You should expect to see increased standard deviation with greater missingness.
3. Compare the parameter estimates and standard errors between the models to initial metrics for assessing the algorithm's performance.

The subsequent steps involve running the MICE algorithm and assessing its performance:

4. For each original dataset, create a duplicate object. Run the mice() function on each copy, setting both m and maxit to 20, seed to 42, and using the default method. This operation should result in three MICE objects.
5. Check convergence of the algorithm by plotting trace plots of the mean and standard deviation, which can be seen when the trace lines "band together" towards a stable value.
6. Fit the linear model to the imputed datasets and pool the results. Compare the parameter estimates and the standard error to the results of the complete case analysis, providing an additional perspective on the MICE algorithm's utility.

Experiment B

1. Similar to A1 (`set.seed(42)` with a 0.15 probability of missingness), generate 1 dataset and Fit the linear model described in the model section to conduct a complete case analysis (Seen in A2)
2. Skipping step A3, create 3 mice objects by using the same `m` and `maxit` values, but vary the method for each one ('pmm', 'sample' and 'rf'). Perform the rest of the steps A5 (checking convergence) and A6 (fitting a linear model, results and compare parameter estimates/standard error) from the above.

Experiment C

1. Instead of generating 3 datasets with 3 probabilities, create 3 datasets using 3 different seeds, (42, 25 and 100) and a constant missingness probability of 0.15. Perform steps A2 (complete case analysis) and A3 (compare parameter estimates for complete case) as before.
2. Similar to A4 (duplicating objects), except set the seed argument to the respective datasets. (E.g. if you are using the dataset with set.seed(42) is used from step one, use seed=42 when creating the mice object). Perform steps A5 (check convergence) and A6 (Compare parameter estimates) as before.

Convergence of the MICE algorithm

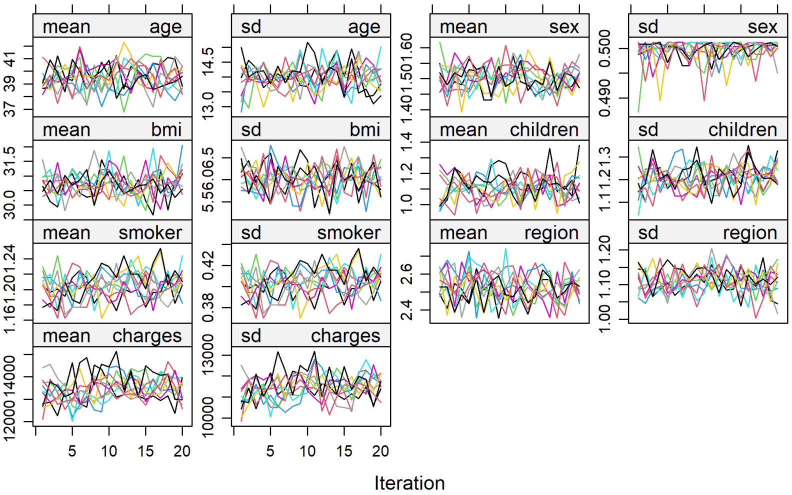

Figure 2: Trace plots for each variable in the insurance dataset with 15% missing data and a seed set to 42. Each line represents a dataset produced by the MICE algorithm with the point for each iteration step. The constant nature of the lines across iterations indicates algorithm convergence.

The constant nature of the lines suggests that after a certain number of iterations, the imputed values have converged around a stable mean and standard deviation. Without checking convergence, it can lead to greater variability of the imputed values in the final pooled dataset, which can bias the linear model parameters.

Experiment A results

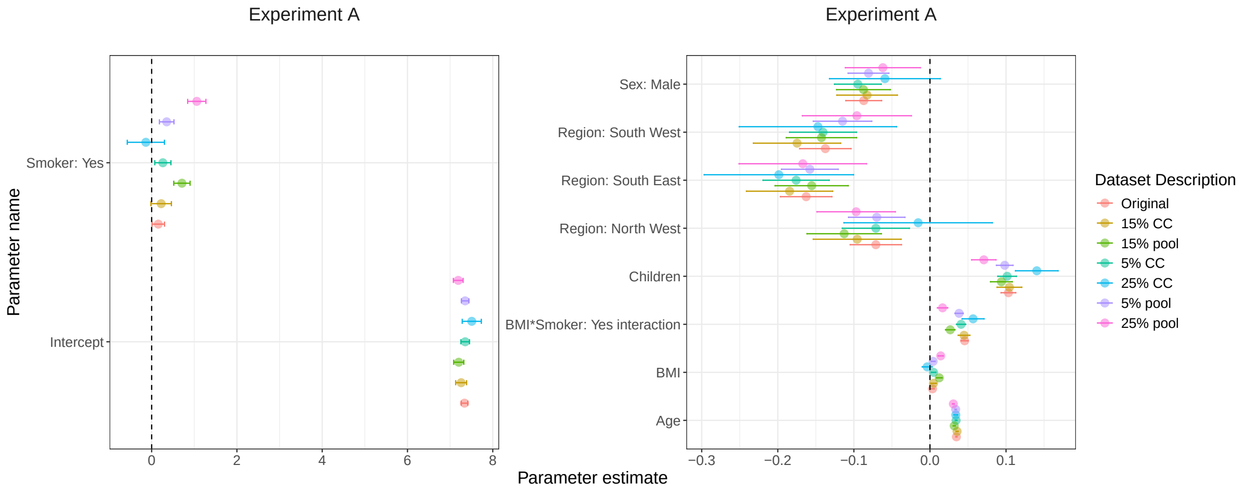

Figure 3: Forest plot against different proprotions of missingness comparing the original, compelte case and imputed datasets. Imputed datasets generally have smaller standard errors than those of complete case datasets. As the proportion of missing data increases, parameter estimates for imputed datasets deviate more from the original, and pooled datasets tend to have standard errors equal to or larger than those of the original. Therefore, increasing missingness will produce more deviation of parameter estimates and the standard error size in pooled datasets compared to the original dataset.

- Imputed datasets have lower standard errors than complete case datasets as the MICE algorithm imputes data into rows with missing values, increasing complete observations.

- Pooled datasets have equal or larger parameter estimate standard errors because while they have the same number of complete cases after imputation, the MICE algorithm standard error also accounts for variance due to missingness, which increases the size of the error bars on these plots.

- While a complete case analysis that yields estimates closer to the original for largely missing datasets might seem more accurate, it doesn't signify superiority over MICE. Greater similarity is due to complete cases depending solely on available data which, while reducing sample size and raising uncertainty, remain within known parameter limits. MICE, however, aims for not replication but the creation of plausible scenarios that factor in the inherent uncertainty of missing values. This results in larger variation in imputed dataset estimates compared to the original, reflecting the natural variability when handling missing values. Thus, although the estimates from imputed datasets might differ from the original dataset, this reflects the algorithm's strength in mimicking real-world uncertainty rather than inaccuracy.

Experiment B Results

- Random Forest excels in handling complex interactions[3], evident in its closer parameter estimates to the original, showing its strength in the context of our dataset's complex relationships like BMI and smoker status. This superior performance over simpler methods is an example of how MICE, and particularly the use of sophisticated algorithms like Random Forest within MICE, can yield more accurate results in complex data scenarios

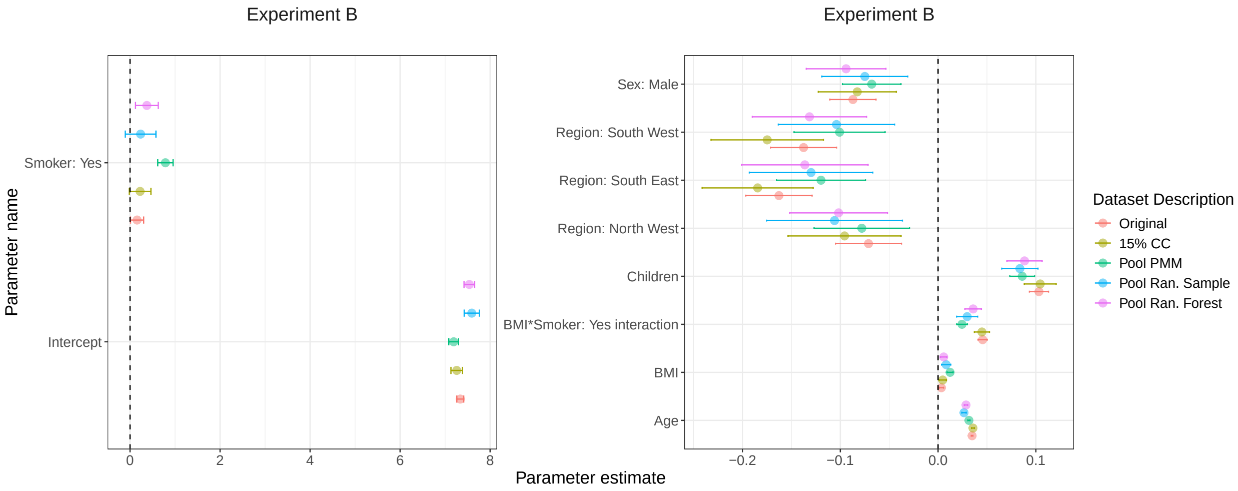

Figure 4: Forest plot of parameter estimates different imputation methods against orgiinal and 15% complete case datasets. PMM has the closest parameter estimates to the 15% complete case dataset and the random forest follows the original data parameter esitmate much more closely than the other imputation methods. Random sample has the largest standard error, suggesting Random Forest imputation method is optimal for our missing dataset.

- It is possible that the PMM pooled dataset has standard errors closer to the original than the expected random forest because PMM is a semi-parametric method, whereas random forest is a non-parametric method. In the case of the MICE algorithm, the non-parametric imputation method like Random Forest does not impose any specific data distribution, possibly leading to a greater risk of noise propagation in the imputed values from the complex model fitting procedure. This increased randomness in the imputed datasets using Random Forest can potentially inflate the standard error estimates in the pooled analysis, compared to those obtained with PMM which employs a more controlled imputation procedure due to its semi-parametric nature[3].

Experiment C results

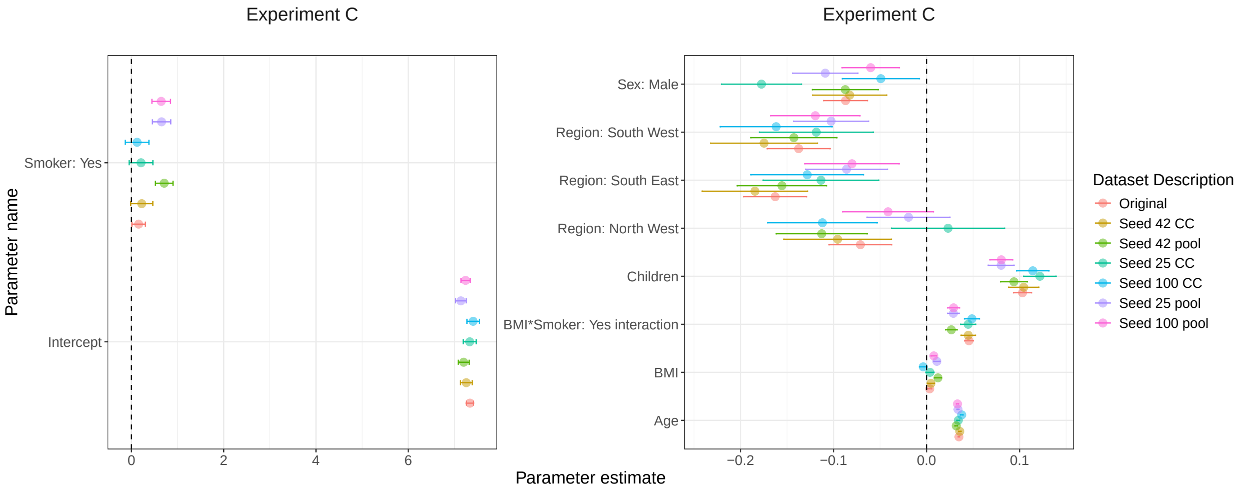

Figure 5: Forest plot of 3 datasets generated using different seeds against the original and complete case. For each seed, the parameter estimates of the complete case against the imputed dataset shifted in has been shifted by approximately the same amount. The parameter estimate and parameter estimate standard error between seed 25, 42 and 100 are similar, suggests that MICE is robust against different distributions of missingness in the dataset. Given the data being MCAR, the MICE algorithm is expected to provide similar imputations across varying missingness distributions, yielding approximately equal standard errors across datasets, as observed in our experiment.

Conclusion

In conclusion, we investigated the effects of missing data on fitting a linear model to a dataset of medical expenses. Our experiment focused on using an algorithm called MICE that helps “fill in” missing data, and we wanted to see how well the algorithm would perform if we altered how much data was missing, changing the way we fill in the missing values and to see if the algorithm could be effective against different versions of the missing data (but similar amount of missing data). It seems the algorithm does well against different missing dataset versions, but can be less reliable when there is a large amounts of missingness. Changing the method of filling the data can be effective, but it depends on the relationship between the variables. In our case, our data had a complex relationship between BMI and smoking status, so a sophisticated filling technique called random forest outperformed others researched in this project.

References

1. Lantz, B. (2013) Machine learning with r. 1st edition. Birmingham: Packt Publishing (Community experience distilled)
2. Dong, Y. and Peng, C.-Y. J. (2013) 'Principled missing data methods for researchers', SpringerPlus, 2(1), pp. 1–17. doi: 10.1186/2193-1801-2-222.
3. Slade, E. and Naylor, M. G. (2020) 'A fair comparison of tree-based and parametric methods in multiple imputation by chained equations', Statistics in medicine, 39(8), pp. 1156–1166. doi: 10.1002/sim.8468.
4. Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice:Multivariate Imputation by Chained Equations in R.Journal of Statistical Software, 45(3), 1-67. DOI10.18637/jss.v045.i03.