How not to compare imputation methods: a warning against using the (root) mean squared error

E P Swens

May 24, 2021

Abstract

This abstract will be added later ...

Introduction

Almost all studies involve missing observations. In psychological and epidemiological clinical research missing data is very frequent (Leurent et al., 2018; Enders, 2017). For instance in randomized studies, patients can be lost to follow-up before the end of the study.

Rubin (1976) introduced three missing data mechanisms; missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). When observations are MCAR, the probability of missingness does not depend on any observed or missing variables. Whereas, MAR occurs when the missingness is conditional on other observed variables. Finally, MNAR infers that the probability of being missing depends on unknown information. Additionally, missingness can be grouped according to a missing data pattern; a matrix describing which values are observed and which values are missing.

To deal with missing values, practitioners rely on the following methods: case deletion, weighting, model-based, and imputation-based procedures. (Schafer, 1997; Little and Rubin, 2019). In this paper, we focus on the latter: imputation.

Imputation is a universal term for filling in missing data with plausible values. Imputation is often a data preprocessing step before the data analysis. It is a favored method because it provides complete data. Therefore, inferences can be obtained over the population of all cases. However, improper imputation could lead to systematic errors of the data analysis estimates. Formally defined as bias: $B(\theta) = E(\hat{\theta}) - \theta$. Common pitfalls of imputation include: omitting important variables from the imputation procedure, dealing with non-normally distributed variables, and plausibility of the missing data mechanism assumption (Sterne et al., 2009).

Surprisingly, we believe another common pitfall is underexposed; evaluating imputation methods. To elaborate, many authors have adapted a method that aims to best recover the true value. This is known mathematically as the (root) mean squared error of the imputed values:

$$rmse = \sqrt{\frac{1}{n_{mis}} \sum_{n=1}^{n_{mis}} (y_i - \dot{y}_i)^2}$$

Where y_i represents the true data value and $\dot{y}_i$ imputed value of the i-th record. For the general case, the minimum RMSE is achieved by predicting the missing y_i by the linear model with the regression weights set to their least squares estimates, otherwise known as regression imputation (Van Buuren, 2018). A disadvantage of this method is that it might still lead to biased parameter estimates, especially with MNAR and MAR mechanisms (Schafer and Graham, 2002). This evidence suggests that the RMSE is a poor estimator for selecting an unbiased imputation method.

In this simulation study, we compared biases and RMSE of different imputation methods. Data was drawn from a multivariate normal distribution. Records were removed based on the MCAR mechanism. Then, the missing values were imputed and followed by data analysis. It was hypothesized that the best RMSE score will not select an unbiased imputation method.

Methods

Results

Conclusion

References

Leurent, B., Gomes, M., & Carpenter, J. R. (2018). Missing data in trial-based cost-effectiveness analysis: An incomplete journey. *Health economics*, *27*(6), 1024–1040.

Enders, C. K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour research and therapy*, *98*, 4–18.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *Bmj*, *338*.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological methods*, *7*(2), 147.