Do We Need Multiple Imputation for Prediction?

Kyle M. Lang

When analyzing incomplete data, we must address the missing values in some way. Missing data imputation is one way in which we can do so. The process of imputing the missing data entails estimating a statistical model and using predictions from this model to replace the missing values. Broadly speaking, we can do missing data imputation in two ways: single imputation (SI) or multiple imputation (MI). When testing hypotheses or doing inference on estimated parameters, SI will produce attenuated standard errors that will inflate Type I error rates and produce overly narrow confidence intervals (CIs; Enders, 2010). MI was developed by Rubin (1978, 1987) to address this limitation and produce accurate statistical inferences from imputed data. MI is much more computationally intensive than SI, however, and there are certain circumstances where the additional computational burden may not be necessary.

Well-implemented SI routines can produce unbiased point estimates of parameters, and there are certain data analytic contexts wherein point estimation is the only objective. Point prediction problems are one such context. Computational efficiency and scalability are often paramount concerns in prediction problems, so SI applications dominate MI in the prediction literature (e.g., García-Laencina, Sancho-Gómez, & Figueiras-Vidal, 2010). Indeed, the unbiasedness of SI suggests that point predictions should also be unbiased when generated from a model fit to singly imputed data. Point predictions are not the whole story, though. In practice, we often want some type of interval estimate around predictions (e.g., a CI or a prediction interval [PI]). As in the case of CIs for model parameters, we would also expect CIs and PIs for predicted values to be too narrow.

In this project, you will use Monte Carlo simulation methods to explore the relative influence of SI and MI on point predictions, CIs for predicted values, and PIs. This project can support up to three students. There are many different ways to parameterize imputation models and many problem characteristics which may affect the relative performance of imputation methods. Therefore, each student will work with the supervisor to define their own operationalization of the problem. Each student will then run an independent simulation study to explore their conceptualization of the problem. Of course, students are encouraged to collaborate on any overlapping portions of their respective codebases to avoid "reinventing the wheel". Throughout this project, we will strive to follow best practices in open science and reproducible research workflows.

References

Enders, C. K. (2010). *Applied missing data analysis*. New York: The Guilford Press.

García-Laencina, P. J., Sancho-Gómez, J.-L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: A review. *Neural Computing & Applications, 19*(2), 263–282. https://doi.org/10.1007/s00521-009-0295-6

Rubin, D. B. (1978). Multiple imputations in sample surveys – A phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association,* 30-34.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (Vol. 519). New York: John Wiley & Sons.

Recommended Reading

For a gentle introduction to the missing data problem and the main ideas of missing data theory, see the sources collected in this reading list.

For a more technically rigorous overview of modern missing data theory, see the sources collected in this reading list.

Required Skills

- Basic R usage
- Linear regression
- Prediction using linear regression
- Statistical inference/Hypothesis testing
- Confidence intervals