Outline & Proposal: Comparing the Performance of Missing Data Methods: Random Forest Imputation, MICE Imputation, Mean Imputation, and Complete Case Analysis for Different Missing Data Mechanisms.

Faizan Khalid Mohsin

October 23, 2020

Contents

1 Overview:

In this paper we will implement and assess the performance of four different missing data methods: complete case analysis, mean imputation, multiple imputation with chained equations (MICE), and Random Forest Imputation (RFI). The goal of this study would be to give a guidance on the performance, in terms of accuracy and computational time, of the different missing data methods for different percentages and types of missing data. We will describe the theoretical frameworks of MICE imputation and Random Forest Imputation and will assess the performance of all four methods at 10%, 20%, 30% and 40% missing data for missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). We will do this for two different simulated data sets. The "missForest" R package by Stekhoven and Bühlmann (2011) will be used for implementing Random Forest Imputation. The papers Strobl et al. (2007), Tang and Ishwaran (2017) and Hong and Lynn (2020) will be used for referring to the theoretical framework. The "MICE" R package by Buuren and Groothuis-Oudshoorn (2010) will be used to implement MICE imputation. The papers Azur et al. (2011), Sterne et al. (2009) and Rubin (1996) will be used for referring to the theoretical framework.

2 Introduction

As data has become more ubiquitous so has the problem of missing data. This is also true for studies which are getting bigger and bigger in scope. To overcome this issue several missing data methods have been established over the years. In this paper will be looking at four missing data methods: complete case analysis, mean imputation, multiple imputation with chained equations, and Random Forest Imputation.

3 Methods

3.1 Data

3.1.1 Types of Missing data

We will describe the three types of missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

3.1.2 Simulating the two data sets

We will create a data set with 1000 observations and 91 variables. For simplicity, all variables will be continuous with 90 variables treated as covariates and one as dependent. The two data sets will have different degree of correlation between the variables.

3.1.3 Inducing missing data

We will introduce missing data using MCAR, MAR, and MNAR at the 10%, 20%, 30% and 40% missingness for each missing data type, hence giving us a total of 12 missing data sets for each of the two data sets.

3.2 Missing Data Methods

3.2.1 Complete Case Analysis

We will describe complete case analysis, its advantages and disadvantages, and cite if its estimates are biased or not.

3.2.2 Mean Imputation

We will describe mean imputation, its advantages and disadvantages, and cite if its estimates are biased or not.

3.2.3 MICE Imputation

We will describe multiple imputation, its theoretical frame work with the key formulas and concepts, and derivations. Specifically, describe multiple imputation with chain equations (MICE) implementation. Also, we will use diagrams to help illustrate the concepts of MICE.

3.2.4 Random Forest Imputation

We will describe random forest imputation, the algorithm, the theoretical framework, the equations, and the formulas. We will use a few helpful diagrams for illustrating the algorithm.

3.3 Assessment of Missing Data Method

To assess the missing data methods we will perform a regression analysis on the imputed data sets and the two complete data sets. The mse's from imputed data sets will be divided by the mse from the corresponding complete data set. We will call this the "standardized mse" and use it to assess the performance of the imputation method benchmarked to the performance of the complete data set.

4 Results

5 Discussion

6 References

Papers to be used for this report.

Azur, Melissa J, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. 2011. "Multiple Imputation by Chained Equations: What Is It and How Does It Work?" *International Journal of Methods in Psychiatric Research* 20 (1). Wiley Online Library: 40–49.

Buuren, S van, and Karin Groothuis-Oudshoorn. 2010. "Mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*. University of California, Los Angeles, 1–68.

Hong, Shangzhi, and Henry S Lynn. 2020. "Accuracy of Random-Forest-Based Imputation of Missing Data in the Presence of Non-Normality, Non-Linearity, and Interaction." *BMC Medical Research Methodology* 20 (1). BioMed Central: 1–12.

Rubin, Donald B. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association* 91 (434). Taylor & Francis Group: 473–89.

Stekhoven, Daniel J., and Peter Bühlmann. 2011. "MissForest—non-parametric missing value imputation for mixed-type data." *Bioinformatics* 28 (1): 112–18. doi:10.1093/bioinformatics/btr597.

Sterne, Jonathan A C, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. 2009. "Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls." *BMJ* 338. BMJ Publishing Group Ltd. doi:10.1136/bmj.b2393.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." *BMC Bioinformatics* 8 (1). BioMed Central: 25.

Tang, Fei, and Hemant Ishwaran. 2017. "Random Forest Missing Data Algorithms." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 10 (6). Wiley Online Library: 363–77.