Statistical Methodology 27 (2015) 82–99

Contents lists available at ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

Navigating choices when applying multiple imputation in the presence of multi-level categorical interaction effects

Aya A. Mitani [a,*], Allison W. Kurian [b,1], Amar K. Das [c], Manisha Desai [d,2]

[a] *Department of Biostatistics, Boston University, 801 Massachusetts Avenue, Boston, MA 02118, United States*

[b] *Department of Medicine and Department of Health Research and Policy, Stanford University School of Medicine, 900 Blake Wilbur, Stanford, CA 94305, United States*

[c] *Department of Psychiatry and The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine, 1 Rope Ferry Rd, Hanover, NH 03755, United States*

[d] *Quantitative Sciences Unit, Department of Medicine, Stanford University School of Medicine, 1070 Arastradero Road, Palo Alto, CA 94304, United States*

ARTICLE INFO

Article history:
Received 18 March 2014
Received in revised form
13 May 2015
Accepted 1 June 2015
Available online 22 June 2015

Keywords:
Multiple imputation
Categorical variables
Interaction effects
Passive imputation
Active imputation

ABSTRACT

Multiple imputation (MI) is an appealing option for handling missing data. When implementing MI, however, users need to make important decisions to obtain estimates with good statistical properties. One such decision involves the choice of imputation model – the joint modeling (JM) versus fully conditional specification (FCS) approach. Another involves the choice of method to handle interactions. These include imputing the interaction term as any other variable (active imputation), or imputing the main effects and then deriving the interaction (passive imputation). Our study investigates the best approach to perform MI in the presence of interaction effects involving two categorical variables. Such effects warrant special attention as they involve multiple correlated parameters that are handled differently under JM and FCS modeling. Through an extensive simulation study, we compared active, passive and an improved passive approach under FCS, as JM precludes passive imputation. We additionally compared JM and

* Corresponding author.
E-mail addresses: amitani@bu.edu (A.A. Mitani), akurian@stanford.edu (A.W. Kurian), amar.das@dartmouth.edu (A.K. Das), manisha.desai@stanford.edu (M. Desai).
1 Tel.: +1 650 498 6004.
2 Tel.: +1 650 725 1946.

http://dx.doi.org/10.1016/j.stamet.2015.06.001

FCS techniques using active imputation. Performance between active and passive imputation was comparable. The improved passive approach proved superior to the other two particularly when the number of parameters corresponding to the interaction was large. JM without rounding and FCS using active imputation were also mostly comparable, with JM outperforming FCS when the number of parameters was large. In a direct comparison of JM active and FCS improved passive, the latter was the clear winner. We recommend improved passive imputation under FCS along with sensitivity analyses to handle multi-level interaction terms.

1. Introduction

Multiple imputation (MI) is an increasingly popular approach for handling missing data [7,8,11,19]. Largely this is due to a growing awareness of the potential bias and inefficiencies that result from applying inappropriate methods, and an increase in software accessibility to perform MI [28]. For example, mainstream software packages such as SAS/STAT software [20], Stata [22], and R [16] offer MI-based analyses. Despite this, a complete-case (CC) analysis, which restricts the analysis to observations with no missing values, still remains the most commonly applied approach, perhaps because it is the default option for handling missing data in all statistical software packages [10,23]. CC analysis, however, is valid when the data are missing completely at random (MCAR) (i.e., missingness is not related to observed or unobserved features), an assumption that does not typically hold in practice. If violated, CC analysis can result in biased and inefficient estimates. MI, on the other hand, is statistically valid under a more flexible assumption about the missing data mechanism; it relies on the assumption that the data are missing at random (MAR) or that missingness is related to observed features only (i.e., after conditioning on relevant observed features, missingness is unrelated to unobserved values). Briefly, MI is a simulation-based approach for filling in each missing datum with a plausible value repeatedly to account for the uncertainty of the sampled values and the imputation process itself. It requires the specification of two statistical models: an imputation model, which is used to impute the missing data for m imputed datasets, and a scientific model, which is used to analyze each of the m imputed datasets in order to address the research question [18].

In addition to being the default approach in software packages, CC analysis may be preferred for its simplicity. Another possible barrier to incorporating MI in the analysis is the numerous choices faced by analysts when implementing MI. Importantly, these choices can have great impact on the results. Among the various choices are the specification of the imputation model (i.e., which variables to consider in the imputation model and their functional form) [5], and the imputation approach. The two main imputation approaches are the joint modeling (JM) approach and the fully conditional specification (FCS) approach. Briefly, JM involves specifying a joint distribution for the data, which is typically assumed to be multivariate normal, in order to derive the posterior predictive distribution from which to impute values [24]. FCS bypasses the specification of a joint model and instead directly specifies the conditional distribution for each partially observed variable [24]. The latter may present advantages for data that contain variables of mixed type, such as binary and categorical variables, where specifying a joint distribution for the data is particularly challenging. While the theoretical properties of estimates generated by JM are well established [14], they are less tractable for FCS, although its use has been well justified empirically through simulation studies [26,29]. In his comprehensive review of these two methods, van Buuren compares and contrasts performance of these two methods. He recommends JM when a multivariate normal assumption is sensible and FCS in the presence of variables of mixed type [24].

Another choice posed to analysts involves how to impute derived variables such as interaction terms. There are two main approaches for handling interaction terms. One approach is to transform