Bull. Aust. Math. Soc. **106** (2022), 349–350
doi:10.1017/S0004972722000673

MULTIPLE IMPUTATION FOR CATEGORICAL VARIABLES IN MULTILEVEL DATA

HELANI DILSHARA KOTTAGE

(Received 15 June 2022; first published online 4 August 2022)

2020 *Mathematics subject classification*: primary 62D10; secondary 62P25.

Keywords and phrases: multiple imputation, missing data, categorical variables, multilevel modelling.

Multiple Imputation (MI) is a technique that imputes a set of plausible values for each missing item using an imputation model. An imputation model predicts a value for the missing item given the observed data and should be compatible with any model that is fitted to the multiply imputed data which is known as the substantive analysis model. To hold the compatibility, the imputation model has to capture the complexities in the substantive analysis model such as the structure of the data set and complex terms (higher order, interaction) considering the type of incomplete variable and the number of incomplete variables in the data set. In this thesis, the application of MI for handling missing values in a set of level-1 categorical variables in a two-level data structure where the data can be fitted with a random intercept substantive analysis model is empirically investigated. Simulation studies considering the missing data rate, the number of clusters and the cluster sizes are based on data generated from a real multilevel educational data set.

In the presence of a set of incomplete categorical variables, this thesis examines the performance of two multivariate multiple imputation techniques: the joint modelling approach which fits a multivariate imputation model for all incomplete variables and the fully conditional specification approach which fits a set of univariate imputation models for incomplete variables. Moreover, two joint modelling approaches are studied: the model that treats incomplete variables as responses and complete variables as predictors, and the model that treats all incomplete and complete variables as responses.

An alternative method of capturing random intercepts in the imputation model is to use fixed effects. The fixed effects imputation model uses a set of dummy variables for representing clusters. The flexibility of the fixed effects approach with multivariate

Thesis submitted to the University of Wollongong in April 2020; degree approved on 16 December 2021; supervisors Carole Birrell and Marijka Batterham.

missingness in categorical variables is assessed with the joint modelling approach and is found to provide poor results.

To represent incomplete categorical variables in the imputation model, latent normal variables are used. The effect of the overcompensation of an ordinal variable with the strategy used for a nominal variable and the impact of under-compensating a nominal variable with the strategy used for an ordinal variable are studied. The findings show that slightly biased results can be expected due to the overcompensation, especially when the sample size is small. The results are heavily biased if a nominal variable is under-compensated.

The semi-parametric multiple imputation method, predictive mean matching, was originally developed for imputing missing values in continuous variables in a single level data structure. In this thesis, for binary and nominal predictors in a random intercept model, the appropriateness of predictive mean matching and the modified algorithm named `midastouch` are studied. The multilevel data structure is accounted for in the imputation phase using either fixed effects or fully observed level-2 predictors. Both the general and modified approaches show promising results for binary variables even with small sample sizes and large missing data rates. However, both approaches of predictive mean matching perform poorly with nominal variables.

When two variables form an interaction term in the substantive analysis model, there is some debate as to the best approach to handle the missing data. Here, an interaction between a binary and a nominal variable in level-1 of a two-level data structure is considered and some drawbacks of using particular methods are identified. The substantive model compatible imputation with the joint modelling approach shows good performance, but in limited circumstances such as with small missing data rates.

Finally, all approaches numerically assessed in the simulation studies are illustrated using a real data analysis example. The application of multiple imputation routines in Progress in International Reading Literacy Study-2016 data for Australia shows some consistent and some contrasting results indicating directions for future research. A set of recommendations are provided to researchers who work with two-level data sets with incomplete categorical predictors in level-1 and fit a random intercept substantive analysis model.

HELANI DILSHARA KOTTAGE, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, New South Wales 2500, Australia
e-mail: hdilk0507@gmail.com

https://doi.org/10.1017/S0004972722000673 Published online by Cambridge University Press