A Bayesian imputation approach for handling missing data in repeatedly measured categorical variables

Oya Kalaycıoğlu[1*+]

[3]Department of Econometrics, Abant Izzet Baysal University, Bolu, Turkey

[*]Corresponding author: oyakalaycioglu@ibu.edu.tr
[+]Speaker: oyakalaycioglu@ibu.edu.tr
Presentation/Paper Type: Poster / Full Paper

Abstract – Multiple imputation (MI) is increasingly being used as a sophisticated tool to handle missing data in the recent years. The standard MI techniques use Bayesian sampling methods for making posterior draws for the imputations and analyse the imputed data sets under a frequentist framework. However, the software packages that are used to implement these techniques are not yet fully flexible to account for the non-normality of the incomplete variables in repeatedly measured data sets. Multiple imputation can be performed in a fully Bayesian modelling context, which offers a flexible alternative in terms of fitting hierarchical non-normal imputation models. This work is motivated by an observational cohort study which consists of different types of incomplete time-varying variables, such as skewed continuous, ordered and unordered categorical variables. These variables were imputed using hierarchical Bayesian imputation models, such as multivariate truncated normal, gamma, multinomial and logistic. The estimates of the analysis model obtained with fully Bayesian imputation method were compared to standard MI methods, which assume normality of the incomplete variables. In addition, the performance of the fully Bayesian imputation was evaluated with a simulation study, under different settings of incomplete data.

Keywords – missing data, multiple imputation, Bayesian imputation, repeated measures, missing at random

I. Introduction

Missingness is a very common problem in repeated measures studies, especially when there are vulnerable patients which are difficult to follow. Multiple imputation (MI) is increasingly being used as a sophisticated tool to handle missing data in the recent years [1]. The standard MI techniques use Bayesian sampling methods for making posterior draws for the imputations and analyse the imputed data sets under a frequentist framework. However, the software packages that are used to implement these techniques are not yet fully flexible to account for the non-normality of the incomplete variables in repeatedly measured data sets. Multiple imputation can be performed in a fully Bayesian modelling context, which offers a flexible alternative in terms of fitting hierarchical non-normal imputation models as well as specifying a correlation structure which conforms to the true correlation of the repeatedly measured data set [2].

In this study, the estimates of the analysis model obtained with fully Bayesian imputation method were compared to standard MI methods, which assume normality of the incomplete variables. In addition, the performance of the fully Bayesian imputation was evaluated with a simulation study, under different settings of incomplete data.

.

II. Motivating data set

The continuity of care data was collected to investigate the association between continuity of cancer care and various health outcomes. It was an observational cohort study with 199 cancer patients, assessed at 5 time points over 12 months. 51% of the time points had at least one incomplete variable, and 15% of the time points had outcome fully observed but at least one incomplete explanatory variable. In Table 1 the variables are described. The study outcome was a continuous variable, with a 39% of missingness. The explanatory variables health need scores are continuous but has skewed distributions. They are repeatedly measured and have significant amount of missingness. The repeatedly measured categorical variables include mental health score and treatment phase.

Table 1. Descriptives of the data set

Variable	Description	Variable type	Mean (SD)	% of missing values
concare	Continuity of care score	Normal	50.6 (9.4)	39.1
phys	Physical need score	Right skewed	10.7 (4.7)	34.6
psyc	Psychological need score	Right skewed	19.6 (8.6)	35.7
hltsys	Health system need score	Right skewed	19.2 (6.7)	35.9
satis	Satisfaction score	Left skewed	41.5 (7.8)	34.7
ghq	Mental health status	Binary	-	36.1
trtphase	Treatment phase	Categorical	-	32.8
site	Cancer site	Categorical	-	0.0

III. Materials and Method

Multiple Imputation replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute, via imputation models. These multiply imputed data sets are then analysed by using standard procedures for complete data and combining the results from these analyses [1].

A. *Multiple Imputation using Multivariate Normal Imputation Models (MVNI)*

MVNI uses a hierarchical multivariate normal imputation model which treats the incomplete variables as multivariate responses of the imputation model [3]. Binary and ordinal variables are imputed assuming a latent normal structure via a probit link which links into the multivariate normal imputation model. For imputing a nominal variable, say X_1 with s categories, s independent latent normal variables are generated according to a probability rule [4]. At each MCMC update a latent normal value is drawn and the missing values in X_1 are imputed using the probability rule.

The analysis model is defined as a random intercepts model:

$concare_{ij} = \beta_0 + \beta_1 phys_{ij} + \beta_2 psyc_{ij} + \beta_3 hltsys_{ij} + \beta_4 satis_{ij} + \beta_5 ghq_{ij} + \beta_6 trtphase_{ij} + \beta_7 lung_{ij} + \beta_8 colorectal_{ij} + \beta_9 period_{ij} + u_i + e_{ij}$,

$e_{ij} \sim N(0, \sigma_e^2)$,

$u_j \sim N(0, \sigma_u^2)$, i=1,...,199 and j=1,…,5.

B. *Imputation by Chained Equations (ICE)*

ICE uses univariate or in other words separate regression models which treats each incomplete variable as response. Software is developed for imputing repeatedly measured continuous variables using random-effects imputation models in R software (Random-effects ICE) [5]. However when there are incomplete binary/categorical variables in dataset, there is no software implementation which fits random-effects logistic imputation models. Therefore fixed effects univariate imputation models are used for imputing repeatedly measured variables. However, using fixed effects imputation models requires an adjustment for repeated measurements.

Therefore the repeated measurements data set is converted to ‘wide’ format. Then separate fixed-effects imputation models for each time point are fitted, and preceding and following time points are used as predictors in the imputation model [6].

C. *Bayesian MI with univariate hierarchical imputation models*

As the standard MI methods have limitations in terms of imputing non-normally distributed repeatedly measured incomplete variables, a fully Bayesian imputation models are proposed as they offer a flexible alternative in terms of fitting hierarchical non-normal imputation models [2,7]. For right skewed variables (i.e. *phys*, *psyc* and *hltsys*) univariate gamma imputation models are used as follows:

$$\mathrm{phys}_{ij} \sim \mathrm{Gamma}\left(\mu_{ij}^{\mathrm{imp}^2}\tau, \mu_{ij}^{\mathrm{imp}}\tau\right)$$

$$\log\left(\mu_{ij}^{\mathrm{imp}}\right) = \gamma_{0i} + \gamma_1 \mathrm{concare}_{ij} + \gamma_2 \mathrm{site{:}\,lung}_i + \gamma_3 \mathrm{site{:}\,col}_i + \gamma_4 \mathrm{period}_{ij}\,,$$

where a $N(\mu_{\gamma_0}^{\mathrm{imp}}, \sigma_{\gamma_0}^{\mathrm{imp}^2})$ prior is specified for the random intercept γ_{0i}. Covariance structure is imposed on the inverse of the variance, i.e. precision parameter τ, with Gamma(0.01, 0.01) prior. Non-informative $N(0, 10^6)$ priors were specified for the regression coefficients, γ_k(k=1,2,3,4) and $\mu_{\gamma_0}^{\mathrm{imp}}$.

For the other skewed variables the following imputation models are used:

- Left skewed variable *satis*: Truncated normal imputation model.
- Binary variable *ghq*: Logistic imputation model
- Categorical variable *trtphase*: Multinomial logistic imputation model

IV. Results

The continuity of cancer care data set is analysed using different MI methods. Data analysis with Bayesian and MVNI methods provided significantly different results. The data set could not be analysed with ICE, due to overfitting problem in the data set.

Table 2. Results from the case study

	Analysis Method			
	Bayesian MI		MVNI	
Variables	$\hat{\beta}_k$	SE($\hat{\beta}_k$)	$\hat{\beta}_k$	SE($\hat{\beta}_k$)
phys	-0.140	0.128	-0.198	0.131
psyc	-0.169	0.041	-0.207	0.049
hltsys	-0.077	0.056	-0.060	0.058
satis	0.498	0.057	0.394	0.068
ghq = 1	-2.047	1.050	-1.619	1.017
trtphase = 1	-3.027	1.273	-1.573	1.506
trtphase = 2	-3.695	1.339	-1.275	1.650
trtphase = 3	-0.222	0.793	0.951	1.819
trtphase = 4	-0.803	1.889	2.362	2.226

In order to evaluate the performance of the fully Bayesian MI in the case of non-normally distributed incomplete repeatedly measured explanatory variables a simulation study is conducted and the results are compared with all available case analysis (AC), MVNI and ICE.

Simulation Strategy: 1000 data sets with 199 subjects were simulated according to:

$$concare_{ij} = -0.100 \times phys_{ij} - 0.175 \times psyc_{ij} - 0.085 \times hltsys_{ij} + 0.400 \times satis_{ij} + 0.200 \times ghq_{ij} + 0.006 \times lung_i + 0.010 \times colorectal_i - 0.100 \times period_{ij} + u_j + e_{ij}$$

with $e_{ij} \sim N(0, 0.7^2)$, $u_j \sim N(0, 0.4^2)$, i=1,...,199 and j=1,…,5

Missing at random (MAR) data were generated with 50% of the time points having missing values. The missingness in each variable was imposed using logistic models that defines probability of missingness.

The simulation study showed that for incomplete binary variables MVNI assuming normality provides biased results, whereas imputing these variables using the univariate Bayesian logistics models as imputation models provided almost unbiased results (Fig. 1). For imputing the skewed incomplete variables, univarite Bayesian non-normal imputaion models provided the least biased results in general. The results of the simulations were also compared in terms of coverage and root mean squred errors of the resgression coefficients of the analysis model. For handling non-normal and categorical incomplete variables the proposed fully Bayesian imputation models were appeared to have the highest coverage and root mean squared error rates compared to standard MI methods and AC analysis.

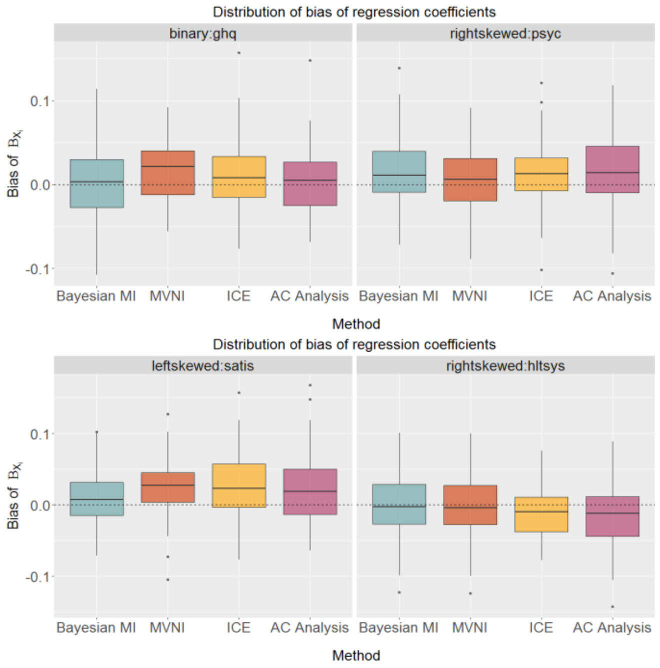

Fig. 1. Box-plots of absolute bias obtained with different methods.

V. CONCLUSION

In this study, a fully Bayesian MI is proposed for handling non-normally distributed incomplete variables. The simulation study showed promising results in the favour of the proposed methodology. That is, the Bayesian MI may be preferable when there is a mixture of incomplete continuous and categorical variables in the repeatedly measured data, particularly for categorical variables. The Bayesian method also offers flexibility regarding the choice of distributions for the continuous variables, which may be useful when handling non-normal continuous variables with missing values. It is important to consider the appropriateness of the chosen distributions carefully, as an incorrect choice may lead to bias.

References

[1] D.B. Rubin, *Multiple imputation for nonresponse in surveys*, New York: John Wiley, 1987.

[2] D. Lunn, N. Best, D. Spiegelhalter, G. Graham and B. Neuenschwander, "Combining MCMC with 'sequential' PKPD modelling," *J Pharmacokinet Pharmacodyn*, vol. 36, pp. 19–38, 2009.

[3] J.L. Schafer, *Analysis of incomplete multivariate data.* London: Chapman and Hall, 1997.

[4] J. Carpenter and M. Kenward, *Multiple imputation and its applications (Statistics in Practice).* West Sussex: John Wiley and Sons, 2013.

[5] S. van Buuren and K. G. Oudshoorn, "MICE: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, vol. 45.

[6] J. Nevalainen, M. G. Kenward and S. M. Virtanen, "Missing values illn longitudinal dietary data: A multiple imputation approach based on a fully conditional specification," *Statistics in Medicine*, vol. 29, pp. 3657-3669, 2009.

[7] G. Carrigan, A. G. Barnett, A. J. Dobson and G. D. Mishra, "Compensating for Missing Data from Longitudinal Studies Using WinBugs," *Journal of Statistical Software*, vol 19, 2007.