Multiple imputation and selection of ordinal level 2 predictors in multilevel models. An analysis of the relationship between student ratings and teacher beliefs and practices

Leonardo Grilli [1], Maria Francesca Marino [1], Omar Paccagnella [2], and Carla Rampichini [1]

[1] Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Florence, Italy

[2] Department of Statistical Sciences, University of Padua, Padua, Italy

---

Address for correspondence: Omar Paccagnella, Department of Statistical Sciences, University of Padua, via Battisti 241, 35121 Padova, Italy.

E-mail: omar.paccagnella@unipd.it.

Phone: (+39) 049 827 4154.

Fax: (+39) 049 827 4170.

Abstract: The paper is motivated by the analysis of the relationship between university student ratings and teacher practices and beliefs, which are measured via a set of binary and ordinal items collected by an innovative survey. The analysis is conducted through a two-level random effect model, where student ratings are nested within teachers. The analysis must face two issues about the items measuring teacher practices and beliefs, which are level 2 predictors: (i) the items are severely affected by missingness due to teacher non-response; (ii) there is redundancy in both the number of items and the number of categories of their measurement scale. We tackle the missing data issue by considering a multiple imputation strategy exploiting information at both student and teacher levels. For the redundancy issue, we rely on regularization techniques for ordinal predictors, also accounting for the multilevel data structure. The proposed solution solves the problem at hand in an original way and it can be applied in general settings requiring to select level 2 predictors affected by missing values. The results obtained with the final model point out that ratings on teacher ability to motivate students are related to certain teacher practices and beliefs.

Key words: Lasso; MICE; missing data; random effects; university course evaluation

1 Introduction

The evaluation of university courses, which is essential for quality insurance, is typically based on student ratings. A large body of literature focuses on studying the factors associated with the expressed evaluations, such as the characteristics of the student, the teacher and the course (Spooren et al., 2013). It is widely recognized that teaching quality is a key determinant of student satisfaction, even if teacher observed characteristics often reveal weak effects (Hanushek and Rivkin, 2006). Therefore, it is helpful to gather more information about teacher practices and beliefs by specific surveys involving the teachers themselves (Goe et al., 2008). In this vein, the PRODID project, launched in 2013 by the University of Padua (Dalla Zuanna et al., 2016), is a valuable source as it implemented a new CAWI survey addressed to the teachers for collecting information on their practices and beliefs about teaching.

We aim at analysing the relationship between student ratings and teacher practices and beliefs, controlling for the available characteristics of students, teachers and courses. Given the hierarchical structure of ratings nested into teachers, we exploit multilevel modelling (Goldstein, 2010; Rampichini et al., 2004). Teacher practices and beliefs from the PRODID survey enter the model as level 2 predictors, but they are missing for nearly half of the teachers due to non-response. Thus, the multilevel analysis must face a serious issue of missing data at level 2. This issue is receiving increasing attention in the literature (Grund et al., 2018). In addition, modelling the effects of teacher practices and beliefs is complicated since they are measured by a wide set of binary and ordinal items, calling for suitable model selection techniques. Therefore, the case study raises the methodological challenge of selecting level 2 predictors affected by missing values. We handle the missing values through multiple imputation by chained equations, exploiting information at both level 1 and level 2 (Grund et al., 2017; Mistler and Enders, 2017). For the selection of predictors, we consider regularization techniques for ordinal predictors (Gertheiss and Tutz, 2010) and we propose a strategy to combine selection of predictors and imputation of their missing values.

The rest of the paper is organized as follows. Section 2 describes the data structure and the model specification. Section 3 outlines the imputation procedure to handle missing data at level 2, then Section 4 presents the regularization method chosen to deal with ordinal predictors. Section 5 outlines the proposed strategy to combine imputation and model selection, while Section 6 illustrates the application of the strategy to the case study, reporting the main results. Section 7 concludes with some remarks and directions for future work.

2 Data description and model specification

As anticipated in Section 1, we wish to analyse the relationship between student ratings and teacher practices and beliefs, controlling for the available characteristics of the student, the teacher, and the course. To this end, we exploit a dataset of the University of Padua for academic year 2012/13, obtained by merging three sources: (i) the traditional course evaluation survey with 18 items on a scale from one to ten, where ten is the maximum; (ii) administrative data on students, teachers, and courses; (iii) the innovative PRODID survey collecting information on teacher practices and beliefs (Dalla Zuanna et al., 2016).

Data have a two-level hierarchical structure, with 56 775 student ratings at level 1 and 1 016 teachers at level 2. The average group size is 79 (min five, max 442).

We investigate student opinion about teacher ability to motivate students, which is one of the items of the course evaluation questionnaire[1]. The analysis is based on the following 2-level linear model for rating i of teacher j:

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\alpha} + \mathbf{z}'_j\boldsymbol{\delta} + \mathbf{q}'_j\boldsymbol{\gamma} + u_j + e_{ij}, \tag{2.1}$$

[1]https:\\www.unipd.it\opinione-studenti-sulle-attivita-didattiche

where $\mathbf{x}_{ij}$ is the vector of level 1 covariates (student characteristics) including the constant, while $\mathbf{z}_j$ is the vector of fully observed level 2 covariates (administrative data on teachers and courses), and $\mathbf{q}_j$ is the vector of partially observed level 2 covariates (teacher practices and beliefs). Model errors are assumed independent across levels with the standard distributional assumptions, namely $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$ and $u_j \overset{\text{iid}}{\sim} N(0, \sigma_u^2)$.

The survey on teacher beliefs and practices has about fifty percent of missing questionnaires, posing a serious issue of missing data at level 2. An analysis based on listwise deletion would discard the entire set of student ratings for non responding teachers, causing two main problems: (*i*) a dramatic reduction of the sample size, and thus of the statistical power, and (*ii*) possibly biased estimates if the missing data mechanism is not MCAR. To overcome these issues, we impute missing values by means of multiple imputation, which allows us to retain all the observations and to perform the analysis under the more plausible MAR assumption (Seaman et al., 2013). In our case, where missing data are nearly 50%, one can have doubts about the opportunity to impute this large amount of missing data. However, simulation results (Marshall et al., 2010) show that MI is better than listwise deletion in terms of bias and coverage of confidence intervals; the performance of MI deteriorates as the percentage of missing values increases, but it is satisfactory up to 50% of missing data.

3 Handling missing data at level 2

In multilevel models, the treatment of missing data requires special techniques since missing values can occur at any level of the hierarchy. Furthermore, missing values can alter variance components and correlations.

Multiple imputation (MI) is a flexible approach to handle missing data taking into account the uncertainty deriving from the imputation procedure. MI is carried out in two steps: (i) generate several imputed data sets according to a suitable imputation model; (ii) fit the substantive model on each imputed data set, and join the results using Rubin rules (Little and Rubin, 2002). The two main approaches to implement MI are *Joint Modelling* (JM) and fully conditional specification, also known as *Multivariate Imputation by Chained Equations* (MICE), see van Buuren (2018) for a comprehensive treatment, while Mistler and Enders (2017) and Grund et al. (2017) for a comparison of these approaches in multilevel settings. In the JM approach, data are assumed to follow a joint multivariate distribution and imputations are generated as draws from the fitted distribution. In the MICE approach, missing data are replaced by iteratively drawing from the fitted conditional distributions of partially observed variables, given the observed and imputed values of the remaining variables in the imputation model. In our case missing data are only at level 2, so we can apply MI techniques to the level 2 data set and then merge level 1 and level 2 data sets. According to the literature on MI in multilevel settings, the imputation model used to simulate missing information at level 2 should include level 2 covariates, the cluster size, and proper summaries of level 1 variables (covariates and response variable). Several strategies may be adopted to summarize level 1 variables: if all the variables are normal, the sample cluster mean is the optimal choice (Carpenter and Kenward, 2013). For the imputation of categorical variables there are no theoretical results, but simulation studies show that the sample cluster mean is a good compromise between accuracy and computational speed (Erler et al., 2016; Grund et al., 2018). Therefore, we summarise level 1 variables through the cluster mean, which is easy to implement in our case since level 1 variables are completely observed.

In our case the imputation step is challenging since we have to impute many categorical variables, indeed about 50% of the teachers did not respond to the whole questionnaire producing missing values on 10 binary items (teacher practices) and 20 ordinal items (teacher beliefs on a seven-point scale). The JM approach, implemented in the R package `jomo` (Quartagno et al., 2019), is computationally demanding, especially in our case with many categorical items. Therefore, we rely on the MICE approach, performing imputations using the `mi chained` command of Stata (Stata Corp., 2017). The imputation model is composed of binary logit models for the 10 binary items (teacher practices) and cumulative logit models for the 20 ordinal items (teacher beliefs). The imputation model includes the following fully observed covariates: teacher characteristics, course characteristics (including the number of ratings), and the cluster means of the ratings for all questions of the course evaluation questionnaire, including the response variable. The inclusion of mean ratings increases the plausibility of the MAR assumption (Grund et al., 2018).

4 Selecting ordinal predictors with regularization techniques

The PRODID questionnaire measures teacher practices using 10 binary items and teacher beliefs using 20 ordinal items on a seven-point Likert scale. Such items contain information on a few dimensions of teaching that in principle could be summarized using latent variable models for ordinal items (Bartholomew et al., 2011). However, about 50% of the teachers did not respond to the questionnaire, thus applying latent variable methods to the complete cases can lead to biased results. On the other hand, fitting latent variable models using imputed data sets raises two main problems: (i) how to combine the results in order to identify the latent dimensions and assign the corresponding scores to the teachers, and (ii) how to take into account the variability of the predicted scores in the main model. The literature on factor analysis in the presence of missing responses is growing (Lorenzo-Seva et al., 2016; Nassiri et al., 2018), but the issue is still controversial, thus we prefer to directly use the imputed PRODID items as covariates in the main model and select them applying model selection techniques.

The imputation method outlined in Section 3 preserves the seven-point scale of the ordinal items. A simple way to specify the effect of an ordinal predictor in a regression model is to treat category codes as scores and include a single regression coefficient. This specification imposes a strong linearity assumption which may be relaxed by dummy coding, where each category is represented by an indicator variable (except for the reference category). For example, fitting model (2.1) with dummy coding for ordinal items $Q12$ and $Q17$ (first category as reference), item $Q12$ does not show a linear effect (Figure 1).

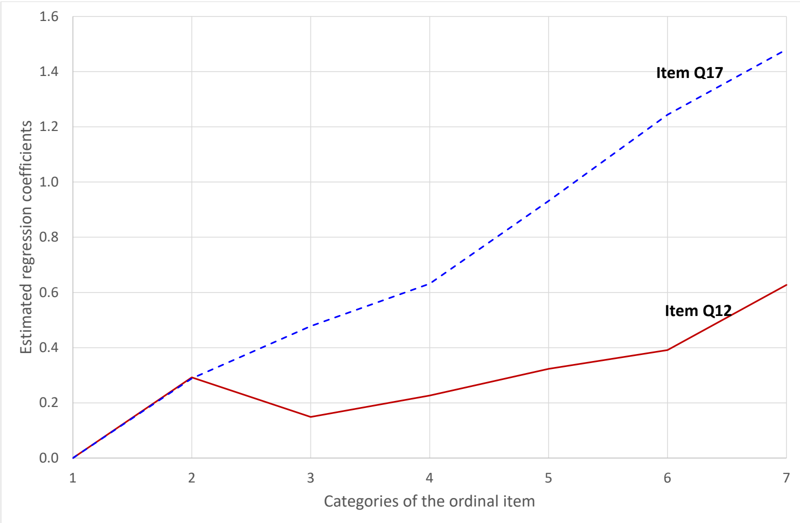

Figure 1: Estimated regression coefficients for items $Q12$ and $Q17$ (dummy coding).

This example makes clear that, in our case, it is advisable to begin with dummy coding for all items, and to apply data driven methods to choose a suitable specification for each item. Indeed, dummy coding yields a flexible but not parsimonious model, since it entails estimating $6 \times 20 = 120$ parameters for the ordinal predictors. Therefore, we need regularization methods allowing us to retain the flexibility of the dummy coding specification, while ensuring model parsimony.

Regularization methods for ordinal predictors (Gertheiss and Tutz, 2010; Tutz and Gertheiss, 2016) have a twofold aim: (i) investigating which variables should be included in the model; (ii) investigating which categories of an ordinal predictor should be distinguished and which can be collapsed. In the presence of K ordinal predictors, each having C_k categories, Gertheiss and Tutz (2010) suggest to implement the *lasso* with the following L_1-penalty term:

$$J(\boldsymbol{\gamma}) = \sum_{k=1}^{K} \sum_{c=2}^{C_k} w_{kc} |\gamma_{kc} - \gamma_{k,c-1}|, \tag{4.1}$$

where γ_{kc} is the coefficient of the dummy variable identifying the c-th category of the kth predictor (with $\gamma_{k1} = 0$ for the baseline category) and w_{kc} are weights allowing for adaptive *lasso*. This approach can be applied to select all items of the PRODID questionnaire, including both ordinal and binary items, since a binary predictor is just an ordinal predictor with $C_k = 2$.

In order to exploit existing software for regularization, we use the *backward difference coding*, also known as *split coding* (Walter et al., 1987; Gertheiss and Tutz, 2010). Such a procedure allows a reparameterisation of the model in equation (2.1) in terms of the new parameters for the ordinal predictors:

$$\tilde{\gamma}_{kc} = \gamma_{kc} - \gamma_{k,c-1} \tag{4.2}$$

and, in turn, the estimation of model parameters by means of a standard *lasso*-type optimisation. Original dummy coefficients are simply obtained by back-transforming $\tilde{\gamma}_{kc}$; that is $\gamma_{kc} = \sum_{r=1}^{c} \tilde{\gamma}_{kr}$. Note that split coding does not affect binary items, so that for such items $\tilde{\gamma} = \gamma$.

The weights w_{kc} in equation (4.1) are chosen as suggested by Zou (2006), yielding an adaptive *lasso* procedure for parameter estimation with the following penalty term:

$$J(\tilde{\gamma}) = \sum_{k=1}^{K} \sum_{c=2}^{C_k} \frac{1}{|\hat{\tilde{\gamma}}_{kc}|} |\tilde{\gamma}_{kc}|, \tag{4.3}$$

with $\hat{\tilde{\gamma}}_{kc}$ denoting the Ordinary Least Squares estimate of model parameter $\tilde{\gamma}_{kc}$.

In practice, in our application, we used the command `lasso2` included in the `lassopack` module of Stata (Ahrens et al., 2018) for parameter estimation. In the following, we outline the regularization algorithm as implemented in this procedure, which relies on Belloni et al. (2012). In particular, the regularization procedure of `lasso2` minimizes the following penalized criterion:

$$Q(\tilde{\gamma}) = \frac{1}{n} RSS(\boldsymbol{\alpha}, \boldsymbol{\delta}, \tilde{\gamma}) + \frac{\lambda}{n} J(\tilde{\gamma}), \tag{4.4}$$

where n is the sample size, $\boldsymbol{\alpha}, \boldsymbol{\delta}$, and $\tilde{\gamma}$ denote the model parameters, $RSS(\boldsymbol{\alpha}, \boldsymbol{\delta}, \tilde{\gamma})$ is the residual sum of squares corresponding to model (2.1) (after split-coding the $\boldsymbol{q}_j$ variables), λ is the overall penalty parameter, and $J(\tilde{\gamma})$ is the penalty term in equation (4.3). To minimize the objective function (4.4), `lasso2` exploits a coordinate descent algorithm (Fu, 1998).

As far as the penalty parameter λ in equation (4.4) is concerned, this is chosen by minimizing the extended BIC index ($EBIC$) proposed by Chen and Chen (2008) and implemented in the `lasso2` procedure as follows:

$$EBIC = n \log(RSS/n) + s \log(n) + 2s \log(p). \tag{4.5}$$

In the equation above, s is the number of parameters in the fitted model, while p is the number of parameters in the full model. Note that $EBIC$ is equal to the standard BIC plus the term $2s \log(p)$.

It is worth to notice that the `lasso2` procedure relies on a standard linear model, while the model of interest appearing in equation (2.1) is a linear random intercept model. We tried a specific procedure for linear mixed models, namely the `lmmlasso` package of `R` (Schelldorfer et al., 2011; Groll and Tutz, 2014), but we encountered computational difficulties due to the large size of the data set. However, the random effects are expected to have a little role in the regularization process for the predictors. Moreover, in order to reduce the bias induced by penalization, it is in general advisable to refit the model using only the selected predictors (Gertheiss and Tutz, 2010; Belloni and Chernozhukov, 2013). Thus, we use the computationally efficient algorithm of `lasso2` to perform variable selection, then we fit the random intercept model (2.1) on the selected predictors.

5 Combining variable selection and multiple imputation

Our case study raises the additional issue of combining variable selection with multiple imputation. Zhao and Long (2017) review different approaches, highlighting that many aspects are still not fully explored. We choose the common approach of applying variable selection on each imputed data set and then combine the results. Specifically, we devise the following strategy:

1. generate M imputed data sets using MICE, as described in Section 3;

2. for each imputed data set, perform variable selection using adaptive lasso for ordinal predictors, as outlined in Section 4;

3. retain the predictors selected in at least $k\%$ of the M imputed data sets; specifically we choose a threshold of 50% as in setting $S2$ of Wood et al. (2008), which performed well in their simulation study;

4. for each imputed data set, fit the linear random intercept model (2.1) including the retained predictors;

5. combine the M vectors of estimated coefficients and the corresponding standard errors exploiting Rubin's rules (Little and Rubin, 2002);
6. perform statistical tests on the regression parameters to refine the set of retained predictors;
7. repeat steps (4)–(6) to choose the final model.

This strategy allows us to select the ordinal predictors while giving proper standard errors, namely accounting for both the hierarchical structure of the data and the uncertainty due to multiple imputation. Step (6) is advisable since it allows us to exploit the availability of proper standard errors to refine variable selection.

6 Results

The strategy outlined in Section 5 is applied to the case study on student ratings presented in Section 2, which raises problems of missing data and selection of ordinal predictors.

The model of interest is the random intercept model (2.1). At level 1, the model includes student predictors $\boldsymbol{x}_{ij}$, which are centered around their cluster average in order to interpret the associated parameters as within effects (Snijders and Bosker, 2012). At level 2, the model includes teacher and course predictors from administrative archives $\boldsymbol{z}_j$ (fully observed), and teacher practices and beliefs $\boldsymbol{q}_j$ (subject to missing). The vector $\boldsymbol{q}_j$ contains dummy variables for 10 binary items and for 20 ordinal items. Adopting the backward-difference coding of Section 4, the total number of parameters for the 20 ordinal items is $6 \times 20 = 120$.

The imputation step is carried out with MICE as described at the end of Section 3. We generate $M = 10$ imputed data sets.

The variable selection procedure begins by applying the regularization method described in Section 4 to each imputed data set, in order to select binary and ordinal items from the PRODID questionnaire, while the other predictors are included in the model without penalization. We retain the predictors selected in at least 50% of the imputed data sets, namely 5 binary items and 13 ordinal items. For each ordinal item k, the procedure selects only a subset of the $\tilde{\gamma}_{kc}$ parameters defined in equation (4.2), implying collapsing of categories. Overall, the regularization procedure reduces the number of parameters $\tilde{\gamma}_{kc}$ from 120 to 26.

The analysis proceeds by fitting model (2.1) with the retained predictors on $M = 10$ imputed data sets and combining the results with the Rubin's rules. The model is fitted by maximum likelihood using the `mixed` and `mi` commands of Stata (Stata Corp., 2017). The variable selection procedure is refined using statistical tests based on the standard errors obtained by Rubin's rules, as suggested in step (6) of Section 5. After this step, the final model includes the binary item $Q02$ and the ordinal items $Q12$, $Q15$, $Q17$, $Q27$. Table 1 reports final model results, while descriptive statistics of model variables are reported in Tables 2 and 3. As shown in Table 1, the selection procedure on ordinal predictors yielded predictor-specific collapsing of categories. For example, for item $Q12$ Table 1 reports two coefficients corresponding to the following collapsed categories: $\{1, 2, 3, 4\}$ (baseline), $\{5, 6\}$, $\{7\}$. This means that the effect of item $Q12$ on the response variable is constant within the collapsed categories. This result is due to the selection procedure, which retained for predictor $Q12$ two out of six parameters in equation (4.2), specifically $\tilde{\gamma}_{12,5}$ and $\tilde{\gamma}_{12,7}$. Due to backward-difference coding, the parameters of the ordinal items represent contrasts between adjacent categories, thus $\hat{\hat{\gamma}}_{12,5} = 0.3204$ is the effect of passing from category $\{1,2,3,4\}$ to category $\{5,6\}$, while $\hat{\hat{\gamma}}_{12,7} = 0.2688$ is the effect of passing from category $\{5,6\}$ to category $\{7\}$. The sum of the two parameters, $0.320\,4 + 0.268\,8 = 0.589\,2$, is the effect of passing from category $\{1,2,3,4\}$ to category $\{7\}$.

We briefly comment on the main finding about the effects of teacher characteristics on their ability to motivate students. Table 1 shows that older teachers and female teachers obtain on average lower ratings, controlling for the remaining covariates. The contribution of external experts ($Q02$) has a positive effect; this is the only item retained by the selection process out of the ten items about practices. As for beliefs, only four out of 20 items are significantly related to the ratings. In particular, ratings tend to be higher for teachers who feel that teaching is an exciting experience ($Q12$) and student opinions are a key indicator of course quality ($Q17$). On the contrary, ratings tend to be lower for teachers who think that cooperation among students helps learning ($Q15$) and teachers interested in discussing didactic methods with colleagues ($Q27$).

In order to assess the overall contribution of teacher practices and beliefs to explain differences in the ratings among courses, we compare the residual level 2 variance under different model specifications. In particular, fitting model (2.1) without any predictor yields an estimated level 2 variance $\hat{\sigma}^2_u = 1.332\,0$, which reduces to $1.230\,6$ (-8%) after introducing all the predictors except teacher practices and beliefs. The final model gives $\hat{\sigma}^2_u = 1.00\,1\,2$, corresponding to a further reduction of residual level 2 variance of about 19%. Thus, teacher practices an beliefs are the most relevant observed factors in explaining differences in the ratings among courses.

Table 1: Multiple imputation estimates: random intercept model for student satisfaction on teacher ability to motivate students.

	Covariates	*Coeff*	*SE*	*p-value*	*FMI†*	*RE†*
Student	Female	-0.0515	0.0188	0.006	0.0000	1.0000
characteristics (lev 1)	Age	0.0479	0.0029	0.000	0.0000	1.0000
	High School grade	0.0072	0.0008	0.000	0.0000	1.0000
	Enrollment year	-0.0918	0.0295	0.002	0.0002	1.0000
	Regular enrollment	-0.1697	0.0485	0.000	0.0000	1.0000
	Passed exams	0.1874	0.0385	0.000	0.0001	1.0000
Course	Compulsory course	-0.2169	0.0431	0.000	.01332	0.9987
characteristics (lev 2)	School					
	Agronomy and Veterinary	-	-	-	-	-
	Social Sciences	0.0516	0.1505	0.732	0.1084	0.9893
	Engineering	-0.3209	0.1279	0.012	0.0785	0.9922
	Psychology	0.2142	0.1619	0.186	0.0526	0.9948
	Sciences	0.0444	0.1340	0.741	0.1662	0.9837
	Humanities	0.2290	0.1393	0.101	0.1544	0.9848
Teacher	Female	-0.1377	0.0793	0.083	0.0987	0.9902
characteristics (lev 2)	Age (years)	-0.0157	0.0039	0.000	0.1266	0.9875
Teacher practices (lev 2)	Q02 External contributors	0.2645	0.0991	0.010	0.4287	0.9589
Teacher beliefs (lev 2)	Q12 Teaching exciting experience					
	{1,2,3,4}	-	-	-	-	-
	{5,6}	0.3204	0.1236	0.012	0.4194	0.9598
	{7}	0.2689	0.0948	0.006	0.3140	0.9696
	Q15 Student cooperation useful					
	{1,2,3,4,5}	-	-	-	-	-
	{6,7}	-0.2338	0.0931	0.015	0.4455	0.9574
	Q17 Student opinions relevant					
	{1,2,3,4}	-	-	-	-	-
	{5}	0.3891	0.1227	0.002	0.4209	0.9596
	{6}	0.3190	0.1308	0.019	0.4890	0.9534
	{7}	0.2340	0.1182	0.050	0.2705	0.9737
	Q27 Discuss teaching methods					
	{1,2}	-	-	-	-	-
	{3,4,5,6,7}	-0.2974	0.1055	0.006	0.3781	0.9636
	Intercept	7.7446	0.2628	0.000	0.1817	0.9822
Residual variances	σ_e^2 (level 1)	3.3971				
	σ_u^2 (level 2)	1.0012				

† *FMI* defined in (6.2), *RE* defined in (6.3) .

Table 2: Descriptive statistics of fully observed variables (56775 ratings, 1016 courses)

	Variables	*Mean*	*sd*	*min*	*max*
Outcome (lev 1)	Student rating on teacher ability	7.387	2.163	1	10
Student	Female	0.511	0.500	0	1
characteristics (lev 1)	Age	20.511	2.962	17	78
	High School grade	80.729	11.817	60	100
	Enrollment year	1.688	0.786	1	3
	Regular enrollment	0.967	0.178	0	1
	Passed exams	0.608	0.259	0	1
Teacher	Female	0.324	0.468	0	1
characteristics (lev 2)	Age (years)	50.638	9.442	32	70
Course	Number of ratings (per course)	79.563	58.902	5	442
characteristics (lev 2)	Compulsory course	0.296	0.457	0	1
	School				
	Agronomy and Veterinary	0.109			
	Social Sciences	0.113			
	Engineering	0.237			
	Psychology	0.075			
	Sciences	0.256			
	Humanities	0.210			

Table 3: Statistics of teacher practices and beliefs (1016 teachers)

Item	*% missing*	*Category*	*Relative Frequency*	
			Observed	*MI average* †
Q02 External contributors	46.75		0.340	0.344
Q12 Teaching exciting experience	47.15			
		1	0.011	0.040
		2	0.017	0.028
		3	0.039	0.046
		4	0.069	0.067
		5	0.229	0.212
		6	0.330	0.307
		7	0.305	0.300
Q15 Student cooperation useful	48.03			
		1	0.019	0.034
		2	0.040	0.052
		3	0.076	0.093
		4	0.142	0.145
		5	0.237	0.214
		6	0.307	0.288
		7	0.180	0.174
Q17 Student opinions relevant	47.24			
		1	0.017	0.024
		2	0.045	0.059
		3	0.065	0.069
		4	0.151	0.159
		5	0.237	0.224
		6	0.289	0.277
		7	0.196	0.189
Q27 Discuss teaching methods	47.83			
		1	0.093	0.127
		2	0.076	0.087
		3	0.113	0.116
		4	0.155	0.139
		5	0.155	0.142
		6	0.226	0.211
		7	0.183	0.178

† MI averages are obtained on 10 data sets imputed by MICE (see Section 3) .

To evaluate the performance of the imputation procedure, the last two columns of Table 1 report the diagnostic measures FMI and RE, which are derived from the decomposition of the total sampling variance V_T of an estimator (Enders, 2010):

$$V_T = V_W + V_B + V_B/M \tag{6.1}$$

where $V_B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}_m - \overline{\hat{\beta}}\right)^2$ is the between-imputation variance, while $V_W = \frac{1}{m}\sum_{m=1}^{M} SE(\hat{\beta}_m)^2$ is the within-imputation variance, with $SE(\hat{\beta}_m)$ denoting the standard error obtained from the m-th imputed data set. The index FMI (Fraction of Missing Information) is used to quantify the influence of multiple imputation on the sampling variance of a parameter estimate:

$$FMI = \frac{V_B + V_B/M}{V_T} \tag{6.2}$$

On the other hand, the index RE (Relative Efficiency) is the relative efficiency for using a finite number of imputations ($M = 10$ in our case) versus the theoretically optimal infinite number of imputations:

$$RE = \left(1 + \frac{FMI}{M}\right)^{-1} \tag{6.3}$$

As regards for level 1 predictors, Table 1 shows values of FMI near zero and values of RE near one. Indeed, level 1 predictors are fully observed and cluster-mean centered, so they are not affected by imputations of level 2 predictors. Fully observed level 2 predictors (i.e. teacher and course characteristics) are little affected by imputations, showing FMI between 0.01 and 0.17, and relative efficiency close to 1. For imputed level 2 predictors (i.e. teacher practices and beliefs) FMI ranges from 0.27 to 0.49, with a mean value of 0.40, indicating that on average 40% of the sampling variance is attributable to missing data, which is lower than the fraction of missing values in the data set (about 50%). This points out a favourable trade-off between the increase of sampling error due to imputations and its reduction due to data augmentation. Moreover, the relative efficiency for imputed predictors ranges from 0.953 to 0.973, suggesting that $M = 10$ imputations are acceptable to obtain a satisfactory level of efficiency.

7 Concluding remarks

In this paper we considered a complex analysis involving a multilevel model with many level 2 ordinal and binary predictors affected by a high rate of missing values. We proposed a strategy to jointly handling missing values and selecting categorical predictors. The proposed solution combines existing methods in an original way to solve the specific problem at hand, but it is generally applicable to settings requiring to select categorical predictors affected by missing values. Specifically, we handled missing data using Multiple Imputation by Chained Equations. This allowed us to retain all the observations and analyze the data under the MAR assumption instead of the unrealistic MCAR assumption. The MAR assumption seems plausible since the imputation model exploits all the information from level 1 and level 2 observed values. The ordinal and binary predictors were selected using an ad hoc regularization method, namely the *lasso* for ordinal predictors. The regularization procedure induces a data-driven specification of the relationship between the response and the ordinal predictors, by collapsing the categories. This method can be easily extend to handle also nominal predictors (Tutz and Gertheiss, 2016). Regularization was applied separately on each imputed data set and the results were combined retaining the parameters selected in at least half of the imputed data sets. Finally, the random effect model of interest was fitted including the chosen predictors. The uncertainty due to imputation is accounted by Rubin's rules. The proposed procedure allowed us to specify the model in a flexible, though parsimonious way.

The results obtained with the final model pointed out that some teacher practices and beliefs are significantly related to ratings about teacher ability to motivate students.

The complexity of the case study, especially in terms of number of observations and number of categorical variables affected by missing values, forced us to rely on computationally low demanding algorithms. The solution of computational issues would allow us to extend the proposed approach in several ways. For example, in the imputation step MICE could be replaced by Joint Modelling (Goldstein et al., 2014; Quartagno and Carpenter, 2016) or by the latent class approach (Vidotto et al., 2018), which give valid inferences for a wider set of specifications of the model of interest, including non linear effects and interactions.

Combining model selection with multiple imputation is an open issue (Zhao and Long, 2017). We devised a simple strategy to face a computationally demanding setting, but it would be interesting to explore other approaches.

Acknowledgements

The authors gratefully acknowledge the support of the University of Padua project SID2016 "Advances in Multilevel and Longitudinal Modelling".

References

Ahrens, A., Hansen, C.B., and Schaffer, M.E. (2018). LASSOPACK: Stata module for lasso, square-root lasso, elastic net, ridge, adaptive lasso estimation and cross-validation. *Statistical Software Components S458458*, Boston College Department of Economics, revised 07 Apr 2018.

Bartholomew, D.J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*, 3rd Edition. John Wiley & Sons Inc.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, **80**, 2369–2429.

Belloni, A., and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, **19**, 521–547.

Carpenter, J., and Kenward, M. (2013). *Multiple imputation and its application.* John Wiley & Sons, Ltd., Chichester, United Kingdom.

Chen, J., and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**, 759–771.

Dalla Zuanna, G., Clerici, R., Paccagnella, O., Paggiaro, A., Martinoia, S., and Pierobon, S. (2016). Evaluative research in education: a survey among professors of University of Padua. *Excellence and Innovation in Learning and Teaching*, **1**, 17–34.

Enders, C.K. (2010). *Applied Missing Data Analysis.* The Guilford Press, New York.

Erler, N.S., Rizopoulos, D., van Rosmalen, J., Jaddoe, V.W.V., Franco, O.H., and Lesaffre, E.M.E.H. (2016). Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian. *Statistics in Medicine*, **35**, 2955–2974.

Fu, W.J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, **7(3)**, 397–416.

Gertheiss, J., and Tutz, G. (2010). Sparse modeling of categorial explanatory variables. *The Annals of Applied Statistics*, **4**, 2150–2180.

Goe, L., Bell, C., and Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis.* National Comprehensive Center for Teacher Quality, Washington, DC.

Goldstein, H. (2010). *Multilevel statistical models*, 4th Edition. John Wiley & Sons, Ltd.

Goldstein, H., Carpenter, J.R., and Browne, W.J. (2014). Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, **177**, 553–564.

Groll, A., and Tutz, G. (2014). Variable selection for generalized linear mixed models by L1-penalized estimation. *Statistics and Computing*, **24**, 137–154.

Grund, S., Ludtke, O., and Robitzsch, A. (2017). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, **21**, 111–149.

Grund, S., Ludtke, O., and Robitzsch, A. (2018). Multiple imputation of missing data at level 2: A comparison of fully conditional and joint modeling in multilevel designs. *Journal of Educational and Behavioral Statistics*, **43**, 316–353.

Hanushek, E.A., and Rivkin, S.G. (2006). Teacher quality. In Hanushek, E.A., and Welch, F. (eds), *Handbook of the economics of education* (Vol. 1), pages 1050–1078. North Holland, Amsterdam.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd Edition. John Wiley & Sons, Inc., Hoboken, New Jersey.

Lorenzo-Seva, U., and Van Ginkel, J.R. (2016). Multiple imputation of missing values in exploratory factor analysis of multidimensional scales: estimating latent trait scores. *Anales de psicología*, **32**, 596–608.

Marshall, A., Altman, D.G., Royston, P., and Holder, R.L. (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Med Res Methodol*, **10**, article 7.

Mistler, S.A., and Enders, C.K. (2017). A comparison of joint model and fully conditional specification imputation for multilevel missing data. *Journal of Educational and Behavioral Statistics*, **42**, 371–404.

Nassiri, V., Lovik, A., Molenberghs, G., and Verbeke, G. (2018). On using multiple imputation for exploratory factor analysis of incomplete data. *Behavior Research Methods*, **50**, 501–551.

Quartagno, M., and Carpenter, J.R. (2016) Multiple imputation of IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine*, **35**, 2938–2954.

Quartagno, M., Grund, S., and Carptenter, J.R. (2019) jomo: A flexible package for two-level Joint Modelling multiple imputation. *The R Journal*, **9(1)**.

Rampichini, C., Grilli, L., and Petrucci, A. (2004). Analysis of university course evaluations: from descriptive measures to multilevel models. *Statistical Methods & Applications*, **13**, 357–373.

Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013). What Is meant by 'Missing at Random'? *Statistical Science*, **28**, 257–268.

Schelldorfer, J., Buhlmann, P., and Van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scandinavian Journal of Statistics*, **38**, 197–214.

Snijders, T.A.B., and Bosker, R.J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*, 2nd Edition. SAGE Publications Inc.

Spooren, P., Brockx, B., and Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, **83**, 598–642.

Stata Corp. (2017). *Stata: Release 15. Statistical Software* . StataCorp LLC, College Station, TX.

Tutz, G., and Gertheiss, J. (2016). Regularized regression for categorical data. *Statistical Modelling*, **16(3)**, 161–200.

van Buuren, S. (2018). *Flexible imputation of missing data*, 2nd Edition. CRC Press Taylor & Francis Group, Boca Raton, FL.

Vidotto, D., Vermunt, J.K., and van Deun, K. (2018). Bayesian multilevel latent class models for the multiple imputation of nested categorical data. *Journal of Educational and Behavioral Statistics*, **43**, 511–539.

Walter, S.D., Feinstein, A.R., and Wells, C.K. (1987). Coding ordinal independent variables in multiple regression analyses. *American Journal of Epidemiology*, **125**, 319–323.

Wood, A.M., White, I.R., and Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, **27**, 3227-3246.

Zhao, Y., and Long, Q. (2017). Variable selection in the presence of missing data: imputation-based methods. *WIREs Comput Stat*, 9:e1402.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101(476)**, 1418–1429.