Statistical Modelling 2022; **22**(**3**): 221–238

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

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

Abstract: The article is motivated by the analysis of the relationship between university student ratings and teacher practices and attitudes, 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 intercept model, where student ratings are nested within teachers. The analysis must face two issues about the items measuring teacher practices and attitudes, which are level 2 predictors: (a) the items are severely affected by missingness due to teacher non-response and (b) 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 addresses the problem at hand in an original way, and it can be applied whenever it is required to select level 2 predictors affected by missing values. The results obtained with the final model indicate that ratings on teacher ability to motivate students are related to certain teacher practices and attitudes.

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

Received November 2019; **revised** June 2020; **accepted** July 2020

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 factors associated with expressed evaluations, including student, teacher and course characteristics (Spooren et al., 2013). It is widely recognized that teaching quality is a key determinant of student satisfaction, even if observed teacher characteristics often reveal weak effects (Hanushek and Rivkin, 2006). Therefore, it is helpful to gather

Address for correspondence: Omar Paccagnella, Department of Statistical Sciences, University of Padua, via Battisti 241, 35121 Padova, Italy.
E-mail: omar.paccagnella@unipd.it

10.1177/1471082X20949710

more information about teacher practices and attitudes 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 Computer-Assisted Web Interviewing (CAWI) survey addressed to teachers for collecting information on their practices and attitudes.

We aim at analysing the relationship between student ratings and teacher practices and attitudes, controlling for available characteristics of students, teachers and courses. Given the hierarchical structure with ratings nested into teachers, we exploit multilevel modelling (Goldstein, 2010; Rampichini et al., 2004). Teacher practices and attitudes 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 attitudes 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 missing values through Multivariate Imputation by Chained Equations (MICE), exploiting information at both levels 1 and 2 (Grund et al., 2017; Mistler and Enders, 2017). For the selection of predictors, we rely on 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 article is organized as follows. Section 2 describes the data and the statistical model. 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. 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 attitudes, controlling for available characteristics at student, course and teacher level. To this end, we exploit a dataset of the University of Padua for academic year 2012/13, obtained by merging three sources: (a) the traditional course evaluation survey with 18 items on a scale from 1 to 10; (b) administrative data on students, teachers and courses; and (c) the innovative PRODID survey collecting information on teacher practices and attitudes (Dalla Zuanna et al., 2016). The dataset used for the analysis is freely available (Felisatti et al., 2020).

The data have a two-level hierarchical structure, with 56 775 student ratings at level 1 and 1 016 teachers at level 2. The median group size is 44 (the 5th percentile is 8; the 95th percentile is 146). Summary statistics of student, teacher and course characteristics are reported in Table 1.

Table 1 Summary statistics of fully observed variables (56 775 ratings, 1 016 courses)

	Variables	*Mean*	SD	Min	Max
Outcome (lev 1)	Student rating on teacher ability	7.387	2.163	1	10
Student characteristics (lev 1)	Female	0.511	0.500	0	1
	Age	20.511	2.962	17	78
	High school grade	80.729	11.817	60	100
	Enrolment year	1.688	0.786	1	3
	Regular enrolment	0.967	0.178	0	1
	Passed exams	6.077	2.587	0	18
Teacher characteristics (lev 2)	Female	0.324	0.468	0	1
	Age (years)	50.638	9.442	32	70
Course 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			

We investigate student opinion about teacher ability to motivate students, which is one of the items of the course evaluation questionnaire (https://www.unipd.it/opinione-studenti-sulle-attivita-didattiche). The analysis is based on the following two-level random intercept linear model for rating i about teacher j:

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

where $\mathbf{x}_{ij}$ is the vector of level 1 covariates (student characteristics) including the constant, $\mathbf{z}_j$ 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 attitudes). Model errors are assumed independent across levels with 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 practices and attitudes has about 50% of missing questionnaires, posing a serious issue of missing data at level 2. Tables 2 and 3 report percentages of missing values and summary statistics for teacher practices and attitudes as observed in the sample, while Figures 1 and 2 show the distributions of ordinal items.

An analysis based on listwise deletion would discard the entire set of student ratings for non-responding teachers, causing two main problems: (a) a dramatic reduction of sample size and statistical power; and (b) possibly biased estimates if the probability of missing observations depends on both model covariates and outcome of interest. To overcome these issues, we impute missing values by means of multiple imputation (MI), which allows us to retain all observations and to perform the analysis under the missing at random (MAR) assumption (Rubin, 1976; Seaman et al., 2013). Given the available covariates, the MAR assumption seems

Table 2 Summary statistics of teacher practices (1= yes, 0= no) for 1 016 teachers

	% miss	*Proportion yes*
Q01 Active learning	46.75	0.850
Q02 External contributors	46.75	0.340
Q03 Student progress monitoring	46.75	0.482
Q04 Integrated evaluation tools	46.75	0.553
Q05 Teaching considering students ratings	46.75	0.821
Q06 Teaching in English lead to change teaching	46.75	0.013
Q07 Teaching supported by multimedia materials	46.75	0.632
Q08 Production of multimedia teaching materials	46.75	0.409
Q09 Advanced use of online platforms	46.75	0.311

Table 3 Summary statistics of teacher attitudes (range 1–7) for 1 016 teachers

	% miss	1st quartile	2nd quartile	3rd quartile
Beliefs				
Q13 Must transmit theoretical knowledge	47.74	3	4	5
Q14 Active teaching stimulate learning	47.24	5	6	6
Q15 Student cooperation useful	48.03	4	5	6
Q16 Advanced technologies promotes student learning	47.64	4	5	6
Q17 Student opinions relevant	47.24	4	5	6
Q18 Single exam better than integrated exam	47.64	2	4	5
Q19 Teacher opinion should be asked	47.93	3	5	6
Q20 Customize teaching according to student needs	47.74	3	5	6
Q21 Teaching in English is an added value	48.62	2	4	6
Needs				
Q22 Make syllabus coherent with learning outcomes	48.13	3	4	5
Q23 Adapt teaching proposal to student training	47.83	4	5	6
Q24 Acquire assessment tools on student learning	47.93	4	5	6
Q25 Consulting teaching experts	47.44	2	4	6
Q26 Training seminars on educational topics	47.44	2	5	6
Q27 Discuss teaching methods	47.83	3	5	6
Q28 Support to integrate technologies in teaching	47.74	2	4	6
Feelings				
Q11 Real passion for teaching	47.15	5	6	7
Q12 Teaching exciting experience	47.15	5	6	7
Q29 Real passion for research	47.24	6	7	7
Q30 Research exciting experience	47.34	6	7	7

plausible in our application. The missingness rate is nearly 50%, thus one can raise doubts about imputing so many 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.

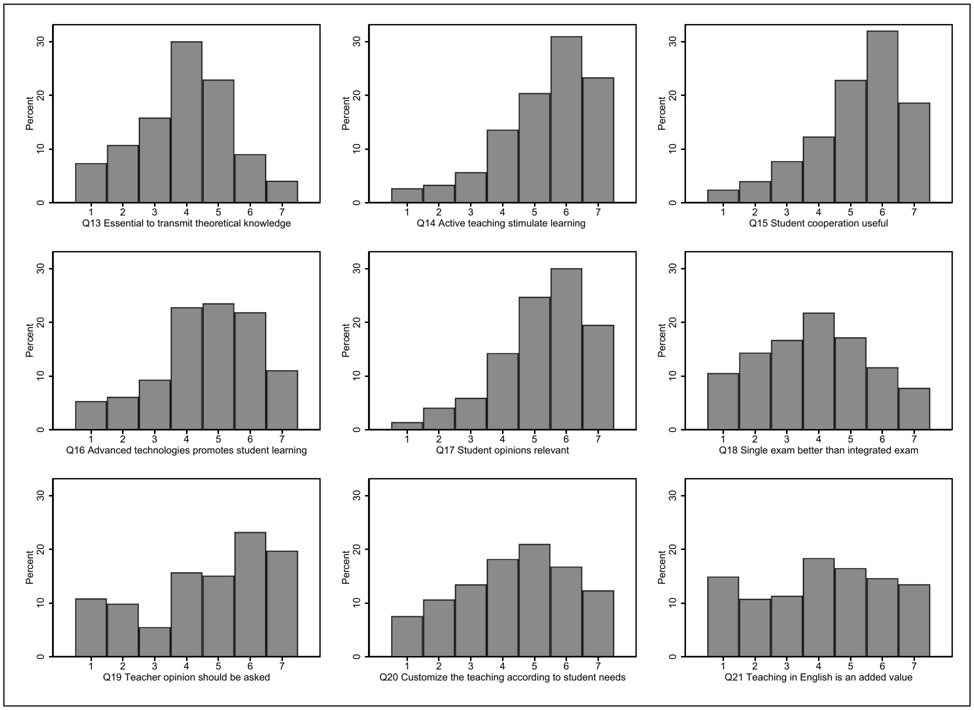

Figure 1 Distributions of items on teacher beliefs

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, if not appropriately handled, missing values can alter variance components and correlations.

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: (a) generate several imputed datasets according to a suitable imputation model; and (b) fit the substantive model on each imputed dataset 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 MICE, see van Buuren (2018) for a comprehensive treatment and 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 imputed 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

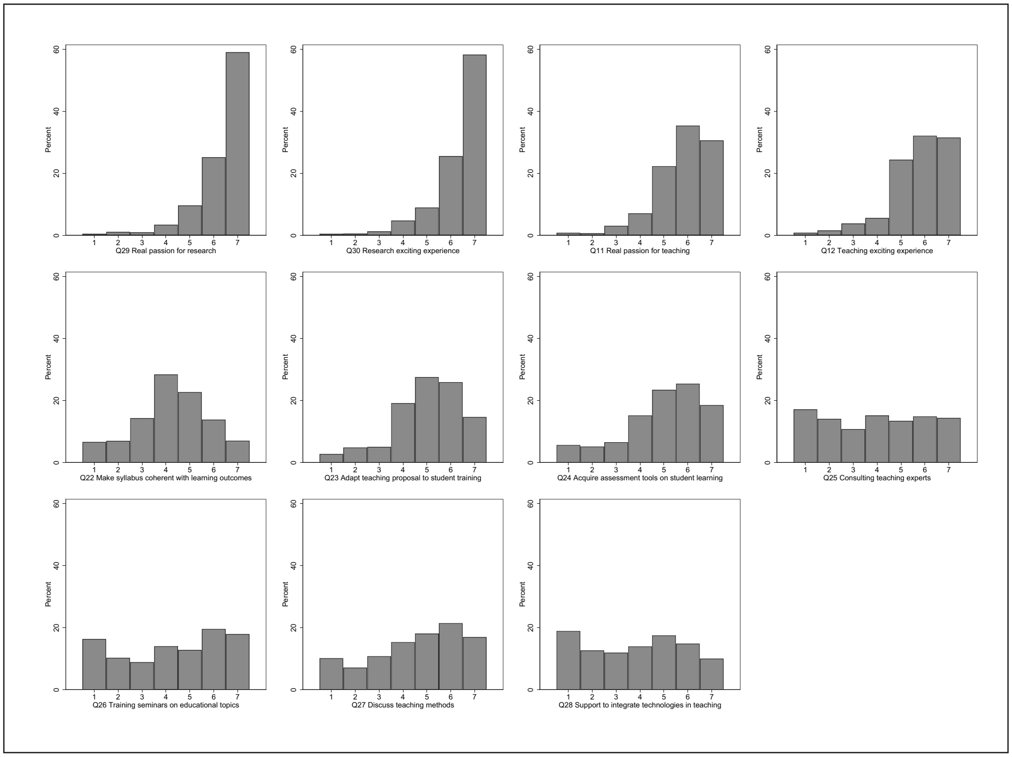

Figure 2 Distributions of items on teacher feelings and needs

only at level 2, so we can apply MI techniques to the level 2 dataset and then merge level 1 and level 2 datasets. According to the literature on MI in multilevel settings (Erler et al., 2016; Grund et al., 2018), the imputation model used to fill in missing information at level 2 should include level 2 covariates, the cluster size and proper summaries of level 1 variables, including the outcome. In our case, level 1 variables helpful for imputation are the ratings of the items of the student questionnaire, even if they are not used as covariates in the analysis model to avoid endogeneity. The response variable and other student ratings are inserted in the imputation model through their sample cluster means, which is shown to be optimal for normal variables (Carpenter and Kenward, 2013). More generally, the cluster mean is a good summary for quantitative covariates, as shown in simulation studies of Erler et al. (2016) and Grund et al. (2018). This approach is easy to implement in our case since imputation is only at level 2.

In our case, the imputation step is challenging: we have to impute many categorical variables since about 50% of teachers did not respond to the whole questionnaire, thus producing missing values on 10 binary items (teacher practices) and 20 ordinal items (teacher attitudes 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 attitudes). The imputation model includes the following types of 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 attitudes using 20 ordinal items on a seven-point Likert scale. Such items bring 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 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 datasets raises two main problems: (a) how to combine the results in order to identify the latent dimensions and assign the corresponding scores to teachers; and (b) how to take into account the variability of 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 simpler way of specifying the effect of an ordinal predictor on the outcome of interest is that of treating category codes as continuous and including a single regression coefficient in the model. However, such a specification relies on a linearity assumption. Furthermore, as highlighted by Gertheiss and Tutz (2009), the interpretation of estimated coefficients is strongly related to the assigned scores which, to some extent, may be arbitrary. To overcome these issues, a dummy coding approach is adopted: each of the K categories of the ordinal predictor is represented by an indicator variable and $K - 1$ coefficients are included in the model. As a result, we obtain a more flexible specification which includes linearity as special case. Clearly, this comes at the cost of an increased number of model parameters and, consequently, a reduction in terms of interpretability. In this respect, we propose using regularization methods that allow 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: (a) investigating which variables should be included in the model; and (b) investigating which categories of an ordinal predictor can be collapsed. For $k = 1, \ldots, K$ 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 cth 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). Specifically, we define a reparameterization of model (2.1) using new parameters for the ordinal predictors

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

which allows us to estimate model parameters by means of a standard *lasso*-type optimization. Then the original parameters are obtained as $\gamma_{kc} = \sum_{r=1}^{c} \tilde{\gamma}_{kr}$. 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 adaptively as suggested by Zou (2006), yielding an adaptive *lasso* procedure for parameter estimation with the following penalty term:

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

with $\hat{\tilde{\gamma}}_{kc}$ denoting the ordinary least squares estimate of $\tilde{\gamma}_{kc}$. As highlighted by Zou (2006), by using adaptive weights we obtain an adaptive *lasso* procedure that enjoys the oracle properties. In detail, it performs as well as if the true underlying model was given in advance; as for the standard lasso approach, the corresponding adaptive version is near-minimax optimal. Lastly, the minimization problem can be solved by the same efficient algorithm for solving the lasso.

In the application, we used the command `lasso2` included in the `lassopack` module of Stata (Ahrens et al., 2020). In the following, we outline the regularization algorithm as implemented in this procedure, which relies on Belloni et al. (2012). In particular, with reference to model (2.1), the regularization procedure of `lasso2` minimizes the penalized criterion

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

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

The penalty parameter λ in equation (4.4) is chosen over a grid of pre-specified values to minimize the extended BIC index (EBIC) proposed by Chen and Chen (2008) and implemented in the `lasso2` procedure as follows:

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

where s and p are the number of parameters of the fitted model and the full model, respectively. Note that EBIC is equal to the standard BIC plus the term $2s\log(p)$. Information criteria such as BIC and EBIC may be preferable to cross-validation in large datasets when the aim is to select the predictors (e.g., Ahrens et al., 2020). Both BIC and EBIC are model-selection consistent if the true model is among the candidate models, but the simulation study of Chen and Chen (2008) shows that EBIC outperforms BIC.

It is worth to note that the `lasso2` procedure relies on a standard linear model, while the model of interest of 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 dataset. 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) using the selected predictors.

5 Combining variable selection and multiple imputation

Our case study raises the additional issue of combining variable selection with MI. While variable selection with fully observed data has been widely investigated, research on this issue for MI datasets is still limited, as underlined by Zhao and Long (2017) and van Buuren (2018). In principle, one could perform variable selection by fitting each candidate model in all imputed datasets and combining the results with Rubin's rules. Variable selection may be performed according to standard techniques such as forward, backward or stepwise search. However, this solution requires intensive computation and it raises issues of overfitting and collinearity (Wood et al., 2008). To overcome these limitations, several alternative solutions have been proposed, which can be divided into three types.

In the first type, variable selection is performed separately on each imputed dataset. With this approach it is likely to obtain different selected variables across imputed datasets. Wood et al. (2008) suggest to retain the covariates according to the so-called *majority rule*, that is, the covariates selected in the majority of imputed datasets. This approach is applied by Shen and Chen (2013) on longitudinal data and by Yang et al. (2005) in a Bayesian framework.

In the second type of solutions, variable selection is performed on a single dataset obtained by stacking all imputed datasets. Weighted regression can be performed on the stacked dataset, applying standard backward selection procedures (Wood et al., 2008) or penalized likelihood using the elastic net penalty (Wan et al., 2015). Chen and Wang (2013) and Marino et al. (2017) use an approach based on a group lasso penalty to guarantee model consistency across different imputations. This approach is extended to longitudinal data by Geronimi and Saporta (2017).

The third type includes solutions that combine variable selection with resampling techniques. Heymans et al. (2007) combine bootstrapping and MI, applying a classical backward selection to each bootstrapped dataset. Musoro et al. (2014) extend this approach adopting a lasso penalization to select variables in each dataset. Long and Johnson (2015) and Liu et al. (2016) combine resampling techniques and MI by using randomized lasso.

In their review, Zhao and Long (2017) support approaches based on lasso, without reaching a clear conclusion about the relative merits of the three types of solutions mentioned earlier.

Using both simulated and real data, Thao and Geskus (2019) propose a comparison between several solutions of variable selection (bootstrap resampling, lasso on original MI datasets and lasso on the stacked dataset) and two different data generating mechanisms under MAR. Their results show that all solutions behave similarly in terms of relative predictive performance and number of retained variables. Therefore, a best approach cannot be identified. Vergouwe et al. (2010) reach the same conclusion comparing the performance of different models with variable selection based on Wald statistics, both on separate and stacked datasets, and the *majority rule* on imputed data.

As the literature does not reach a consensus on the optimal solution, in the light of the complexity of our application, we propose a mixed solution easy to implement and computationally convenient. First, we perform variable selection on each imputed dataset using lasso, and we specify a provisional model using the *majority rule*. Then, we fit the provisional model on each imputed dataset and we combine the results using Rubin's rules in order to refine variable selection with statistical tests. Specifically, we propose the following strategy:

1. Generate M imputed datasets using MICE, as described in Section 3;
2. For each imputed dataset, perform variable selection using adaptive lasso for ordinal predictors, as outlined in Section 4;
3. Retain the predictors selected in at least $k\%$ of M imputed datasets; specifically, we apply the *majority rule* ($k = 50\%$) as suggested by Wood et al. (2008);
4. For each imputed dataset, 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, in particular, we perform Wald tests using the combined standard errors, retaining the predictors significant at level $\alpha = 0.10$;
7. Repeat steps (4)–(6) until only statistically significant predictors remain.

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 MI. Step (6) is advisable since it allows us to exploit 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 $\boldsymbol{x}_{ij}$ (see Table 1), which are centred 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, see Table 1), and teacher practices and attitudes $\boldsymbol{q}_j$ (subject to missing). The vector $\boldsymbol{q}_j$ contains dummy variables for 10 binary items and for 20 ordinal items (see Tables 2 and 3). 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 datasets. In most applications, $M = 10$ is large enough to obtain efficient estimators. We establish that $M = 10$ is adequate in our application on the basis of the Relative Efficiency (RE) index (see later Table 4).

The variable selection procedure begins by applying the regularization method described in Section 4 to each imputed dataset, 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 imputed datasets, 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 datasets 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 *Q*02 and the ordinal items *Q*12, *Q*15, *Q*17 and *Q*27. Table 4 reports the results of the final model.

Indeed the selection procedure on ordinal predictors yielded predictor-specific collapsing of categories. For example, for item *Q*12, Table 4 reports two coefficients corresponding to the following collapsed categories: $\{1, 2, 3, 4\}$ (baseline), $\{5, 6\}$ and $\{7\}$. This means that the effect of item *Q*12 on the response variable is constant within the collapsed categories. This result is due to the selection procedure, which retained for predictor *Q*12 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 $\hat{\tilde{\gamma}}_{12,5} = 0.3204$ is the effect of passing from category $\{1, 2, 3, 4\}$ to category $\{5, 6\}$, while $\hat{\tilde{\gamma}}_{12,7} = 0.2688$ is the effect of passing from category $\{5, 6\}$ to category $\{7\}$. The sum of the two parameters, $0.3204 + 0.2688 = 0.5892$, is the effect of passing from category $\{1, 2, 3, 4\}$ to category $\{7\}$.

Student and course characteristics are inserted in the model as control variables, thus we do not comment their effects. As for teacher characteristics, we note that older teachers and female teachers obtain on average lower ratings on the ability to motivate students, controlling for the remaining covariates. As for practices, the contribution of external experts (*Q*02) has a positive effect; this is the only item retained by the selection process out of the 10 items about practices. As for attitudes, only 4 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 (*Q*12) and teachers who believe that student opinions are a key indicator of course quality (*Q*17). On the contrary, ratings tend to be lower for teachers who think that cooperation among students helps learning (*Q*15) and teachers interested in discussing didactic methods with colleagues (*Q*27).

In order to assess the overall contribution of teacher practices and attitudes in explaining 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 $\hat{\sigma}_u^2 = 1.3320$, which reduces to 1.2306 (−8%) after introducing all predictors except teacher practices and attitudes. The final model gives $\hat{\sigma}_u^2 = 1.0012$, corresponding to a further reduction of about 19%. Thus, teacher practices and attitudes are the most relevant observed factors in explaining differences in the ratings among courses.

To evaluate the performance of the imputation procedure, the last two columns of Table 4 report the diagnostic measures FMI and RE, which are derived from the decomposition of the total sampling variance V_T of an estimator (e.g., 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 dataset. The index Fraction of Missing Information (FMI) is used to quantify the influence of MI on the sampling variance

Table 4 Multiple imputation estimates: Random intercept model for student satisfaction on teacher ability to motivate students

	Covariates	*Coeff*	*SE*	*P-value*	*FMI†*	*RE†*
Student characteristics (lev 1)	Female	–0.0515	0.0188	0.006	0.0000	1.0000
	Age	0.0479	0.0029	0.000	0.0000	1.0000
	High school grade	0.0072	0.0008	0.000	0.0000	1.0000
	Enrolment year	–0.0918	0.0295	0.002	0.0002	1.0000
	Regular enrolment	–0.1697	0.0485	0.000	0.0000	1.0000
	Passed exams	0.1874	0.0385	0.000	0.0001	1.0000
Course characteristics (lev 2)	Compulsory course	–0.2169	0.0431	0.000	.01332	0.9987
	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 characteristics (lev 2)	Female	–0.1377	0.0793	0.083	0.0987	0.9902
	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 attitudes (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				

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

of a parameter estimate:

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

On the other hand, the index RE is the RE 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 for level 1 predictors, Table 4 shows values of FMI near zero and values of RE near one. Indeed, level 1 predictors are fully observed and cluster-mean centred, 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 RE close to 1. For imputed level 2 predictors (i.e., teacher practices and attitudes), 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 dataset (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 RE for imputed predictors ranges from 0.953 to 0.973, suggesting that $M = 10$ imputations ensure a satisfactory level of efficiency.

7 Concluding remarks

In this article, 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 handle 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. Since missing data are only at level 2, the imputation model is based on level 2 units, while exploiting level 1 information through cluster means of level 1 variables. Specifically, we handled missing data using MICE. This allowed us to retain all observations, thus obtaining more efficient estimates with respect to a complete case analysis. The MAR assumption underlying MI seems plausible given the wealth of information in levels 1 and 2 observed values exploited by the imputation model. 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 extended to handle also nominal predictors (Tutz and Gertheiss, 2016). Our solution is a novelty in the limited literature on variable selection under MI, where the focus is mainly on continuous and binary covariates (Thao and Geskus, 2019). Like in Wood et al. (2008), regularization was then applied separately on each imputed dataset and results were combined retaining the parameters selected in at least half of imputed datasets. 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, which is especially important in a multilevel framework.

The results obtained with the final model pointed out that some teacher practices and attitudes 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, suggested to carry out imputations using MICE, which is computationally low demanding. The solution of computational issues would allow us to explore the performance of other approaches for the imputation step, such as JM (Goldstein et al., 2014; Quartagno and Carpenter, 2016) or the latent class approach (Vidotto et al., 2018). Such methods implement more general imputation models, thus allowing a wider set of specifications of the analysis model, including non-linear effects, interactions and/or random slopes. This greater flexibility is especially important to achieve congeniality (van Buuren, 2018) when the imputation step is performed without taking into account the specification of the analysis model.

Combining model selection with MI is an open issue (Zhao and Long, 2017). We devised a simple strategy to face a computationally demanding setting, following the approach of selecting the variables in each imputed dataset and pooling the results. It would be interesting to explore other approaches, also through simulation studies, such as variable selection on a stacked imputed dataset and solutions that combine variable selection with resampling techniques.

Supplementary materials

The Stata code for reproducing the analysis is available from the journal's repository http://www.statmod.org/smij/archive.html.

Acknowledgements

The authors thank three anonymous referees for their helpful comments and suggestions.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

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

References

Ahrens A, Hansen CB and Schaffer ME (2020) lassopack: Model selection and prediction with regularized regression in Stata. *The Stata Journal*, **20**, 176–235.

Bartholomew DJ, Knott M and Moustaki I (2011) *Latent Variable Models and Factor Analysis: A Unified Approach, 3rd edition*. Hoboken, NJ: 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. it Bernoulli, **19**, 521–47.

Carpenter J and Kenward M (2013). *Multiple Imputation and Its Application*. Chichester: John Wiley & Sons Ltd.

Chen J and Chen Z (2008) Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**, 759–71.

Chen Q and Wang S (2013) Variable selection for multiply-imputed data with application to dioxin exposure study. it Statistics in Medicine, **32**, 3646–59.

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 CK (2010) *Applied Missing Data Analysis*. New York: The Guilford Press.

Erler NS, Rizopoulos D, van Rosmalen J, Jaddoe VWV, Franco OH and Lesaffre EMEH (2016) Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian. *Statistics in Medicine*, **35**, 2955–74.

Felisatti E, Dalla Zuanna G, Serbati A, Clerici R, Paggiaro A, Martinoia S, Stocco C, Pierobon S, Aquario D, Da Re L and Paccagnella O (2020) *PRODID project. Dataset*. Research Data Unipd. doi: 10.25430/researchdata.cab.unipd.it.00000359

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

Geronimi J and Saporta G (2017) Variable selection for multiply-imputed data with penalized generalized estimating equations. *Computational Statistics and Data Analysis*, **110**, 103–14.

Gertheiss J and Tutz G (2009) Penalized regression with ordinal predictors. *International Statistical Review*, **77**, 345–65.

Gertheiss J and Tutz G (2010) Sparse modeling of categorial explanatory variables. *The Annals of Applied Statistics*, **4**, 2150–80.

Goe L, Bell C and Little O (2008) *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Goldstein H (2010) *Multilevel Statistical Models, 4th edition*. John Wiley & Sons Ltd.

Goldstein H, Carpenter JR and Browne WJ (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**, 553–64.

Groll A and Tutz G (2014) Variable selection for generalized linear mixed models by L1-penalized estimation. *Statistics and Computing*, **24**, 137–54.

Grund S, Ludtke O and Robitzsch A (2017) Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, **21**, 111–49.

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**, 316–53.

Hanushek EA and Rivkin SG (2006) Teacher quality. In *Handbook of the Economics of Education*, edited by EA Hanushek and F Welch, Vol. 2, pages 1050–1078. North Holland, Amsterdam: Elsevier.

Heymans MW, van Buuren S, Knol DL, van Mechelen W and de Vet HCW (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Medical Research Methodology*, **7**, doi: 10.1186/1471-2288-7-33.

Heymans MW, van Buuren S, Knol DL, van Mechelen W and de Vet HCW (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Medical Research Methodology*, **7**, doi: 10.1186/1471-2288-7-33.

Liu Y, Wang Y, Feng Y and Wall MM (2016) Variable selection and prediction with incomplete high-dimensional data. *The Annals of Applied Statistics*, **10**, 418–50.

Long Q and Johnson BA (2015) Variable selection in the presence of missing data: Resampling and imputation. *Biostatistics*, **16**, 596–610.

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

Marino M, Buxton OM and Li Y (2017) Covariate selection for multilevel models with missing data. *Stat*, 6, 31–46.

Marshall A, Altman DG, Royston P and Holder RL (2010) Comparison of techniques for handling missing covariate data within prognostic modelling studies: A simulation study. *BMC Medical Research Methodology*, **10**, article 7.

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

Musoro JZ, Zwinderman AH, Puhan MA, ter Riet G and Geskus RB (2014) Validation of prediction models based on lasso regression with multiply imputed data. *BMC Medical Research Methodology*, **14**. doi: 10.1186/1471-2288-14-116.

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**, 501–51.

Quartagno M and Carpenter JR (2016) Multiple imputation of IPD metaanalysis: Allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine*, **35**, 2938–54.

Quartagno M, Grund S and Carptenter JR (2019) jomo: A flexible package for two-level joint modelling multiple imputation. *The R Journal*, **9**. URL https://journal.r-project.org/archive/2019/RJ-2019-034/RJ-2019-034.pdf (last accessed 22 September 2020).

Rampichini C, Grilli L and Petrucci A (2004) Analysis of university course evaluations: from descriptive measures to multilevel models. *Statistical Methods & Applications*, **13**, 357–73.

Rubin DB (1976) Inference and missing data. *Biometrika*, **63**(**3**) 581–92.

Seaman S, Galati J, Jackson D and Carlin J (2013) What is meant by 'missing at random'? *Statistical Science*, **28**, 257–68.

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.

Shen C-W and Chen Y-H (2013) Model selection of generalized estimating equations with multiply imputed longitudinal data. *Biometrical Journal*, **55**, 899–911.

Snijders TAB and Bosker RJ (2012) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, 2nd edition*. London: SAGE Publications.

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*. College Station, TX: StataCorp LLC.

Thao LTP and Geskus R (2019) A comparison of model selection methods for prediction in the presence of multiply imputed data. *Biometrical Journal*, **61**, 343–56.

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*. Boca Raton, FL: CRC Press Taylor & Francis Group.

Vergouwe Y, Royston P, Moons KGM and Altman DG (2010) Development and validation of a prediction model with missing predictor data: A practical approach. *Journal of Clinical Epidemiology*, **63**, 205–14.

Vidotto D, Vermunt JK 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**, 511–39.

Walter SD, Feinstein AR and Wells CK (1987) Coding ordinal independent variables in multiple regression analyses. *American Journal of Epidemiology*, **125**, 319–23.

Wan Y, Datta S, Conklin DJ and Kong M (2015) Variable selection models based on multiple imputation with an application for predicting median effective dose and maximum effect. *Journal of Statistical Computation and Simulation*, **85**, 1902–16.

Wood AM, White IR and Royston P (2008) How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, **27**, 3227–46.

Yang X, Belin TR and Boscardin WJ (2005) Imputation and variable selection in linear regression models with missing covariates. *Biometrics*, **61**, 498–506.

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**, 1418–29.