12th International Multilevel Conference
Utrecht, 9-10 April 2019

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

Leonardo Grilli[1], M. Francesca Marino[1],
Omar Paccagnella[2], Carla Rampichini[1]

[1] Department of Statistics, Computer Science, Applications – University of Florence
[2] Department of Statistical Sciences – University of Padua

Motivation

We wish to analyse the relationship between

student satisfaction about teaching

and several **student and teacher characteristics**

Methodological issues:

- The data have a **multilevel structure** (student ratings nested into teachers) → multilevel modelling
- Teacher practices and beliefs are **missing** for about 50% of the teachers → multiple imputation (MI) of level 2 predictors
- Teacher practices and beliefs are measured by **many binary/ordinal items** → selection of ordinal predictors
- Key issue: combine multiple imputation and selection of ordinal predictors

Data (merging 3 sources)

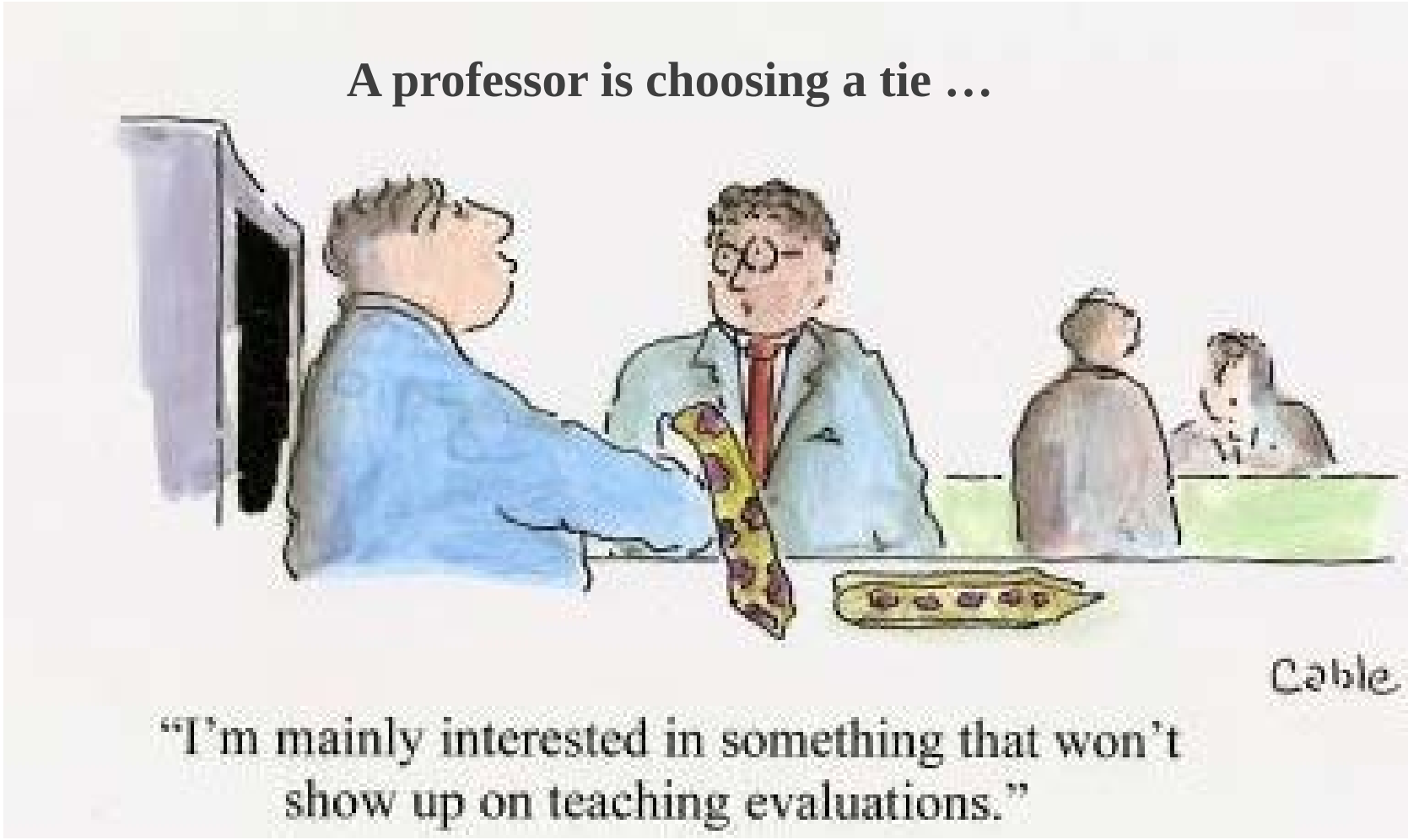
University of Padua, 3-year degree programs, academic year 2012/13

1. **Student ratings** about the course: 18 items on a 10-point scale
2. **Administrative data** on characteristics of students (age, gender, exams, …), teachers (age, gender, …) and course (hours, compulsory,…)
3. **Web survey on teachers (PRODID survey)** to collect data on ***practices*** (10 binary items) and ***beliefs*** (20 ordinal items, scale 1 to 7) – Dalla Zuanna et al., Technical Report Series, n. 1, September 2014

Multilevel structure

- Level 1: student ratings (n=56775)
- Level 2: teachers (n=1016)

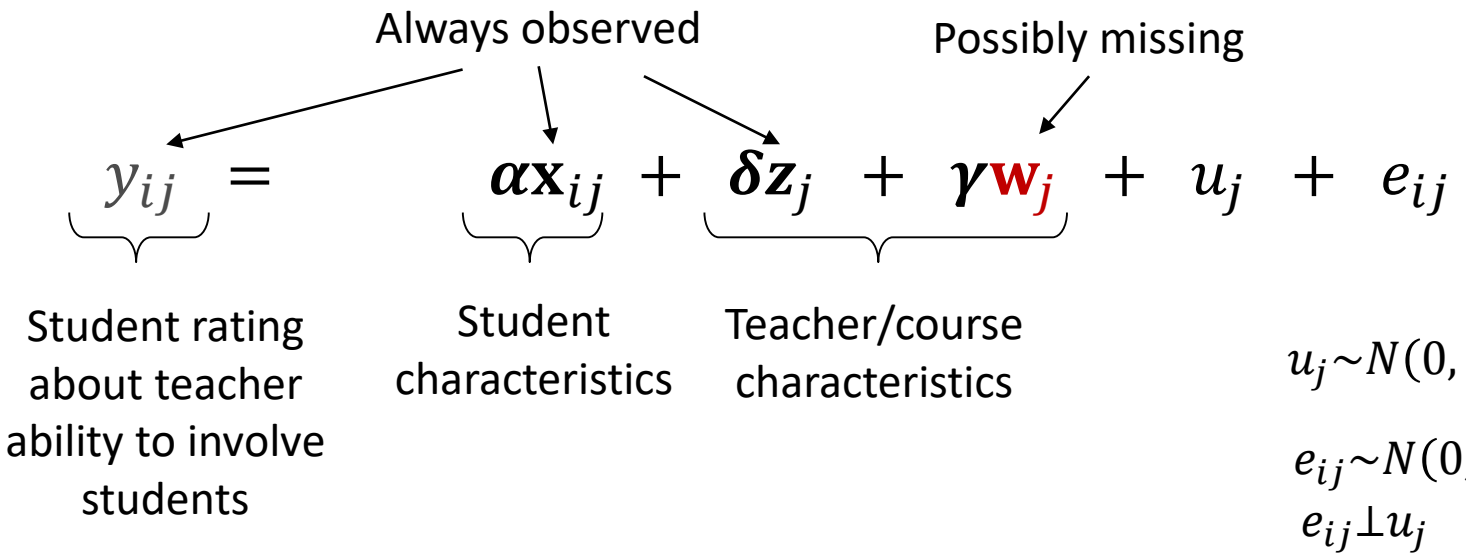
A professor is choosing a tie …



The substantive model

The analysis focuses on **teacher ability to involve students** (item D06)

- We specify a ***random intercept linear model***
  - Level 1: student rating (index *i*)
  - Level 2: teacher (index *j*)



Nonresponse on teacher practices and beliefs

The non-response rate was nearly 50%

→ missing values on level 2 predictors ***w***

How to tackle this missing data problem?

course	Y	X	W
1	$y_{11} \dots y_{1n_1}$	$x_{11} \dots x_{1n_1}$	w_1
2	$y_{21} \dots y_{2n_1}$	$x_{21} \dots x_{2n_1}$	w_2
…	…	…	…
j	$y_{j1} \dots y_{jn_j}$	$x_{j1} \dots x_{jn_j}$	*MISSING*
…	…	…	…
J	$y_{J1} \dots y_{Jn_j}$	$x_{J1} \dots x_{Jn_j}$	w_J

Naive approach: list wise deletion

if a level 2 predictor w_j is missing → delete the entire course, i.e. all the ratings at level 1

Handling missing data

- List wise deletion is not recommended: it is extremely inefficient (deleting about 50% of the observations) and inferences are valid only under the unrealistic MCAR assumption
- **Multiple Imputation (MI)** is a flexible approach retaining all observations and based on MAR (i.e. the probability of a value being missing can depend on observed values)
- The substantive model is a random intercept model, however our imputation problem is similar to standard MI, since imputation involves only level 2 predictors so we proceed as follows:
  i. impute missing values on the dataset of level 2 units (*M* imputations)
  ii. merge level 1 and level 2 datasets to obtain *M* 'complete' datasets

Imputation model at level 2: ingredients

- The imputation model at level 2 should include
  - all level 2 predictors
  - summaries of level 1 variables (level 1 predictors + response variable)
- *Which is the appropriate summary of a level 1 variable for the imputation model at level 2?*
  - Continuous variable in a normal linear model → the **sample cluster mean** is optimal (Carpenter and Kenward, 2013)
  - Categorical variable → no theoretical results, but in the simulation studies the **sample cluster mean** seems to be a good compromise between accuracy and computational speed (Erler et al., 2016, Statistics in Medicine; Grund et al., 2017, J. Educ. Behavioral Statistics).

Imputation model at level 2: specification

- We adopt Fully Conditional Specification (FCS), also known as Multiple Imputation by Chained Equations (MICE) (van Buuren & Oudshoorn, 1999), implemented for example in the R package mice and in the mi procedure of Stata.
- For the q-th level 2 variable affected by missing values (q=1, …, 30), imputations are generated from the fully conditional distribution

$$P\left(w_{jq} \mid \underbrace{\bar{\mathbf{y}}_j, \bar{\mathbf{x}}_j,}_{\text{Cluster means of level 1 variables (always observed)}} \underbrace{\mathbf{z}_j,}_{\text{Level 2 predictors (always observed)}} \underbrace{\mathbf{w}_{j(-q)}}_{\text{Currently imputed values of other level 2 predictors affected by missing values}} \right)$$

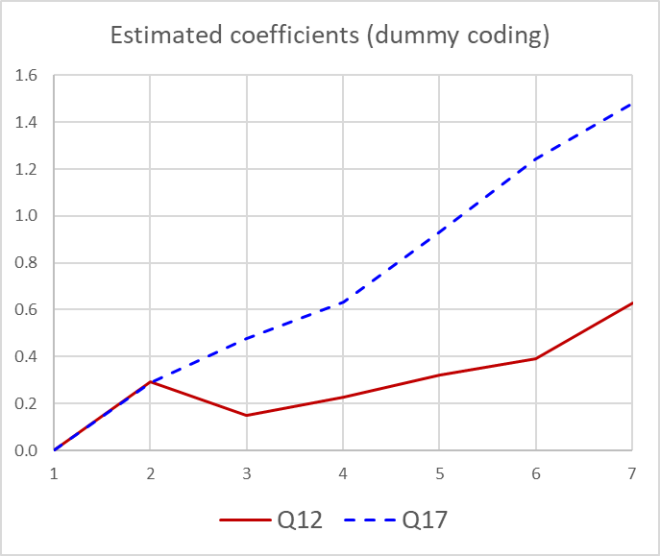
  - The 30 conditional distributions are modelled via GLMs (binary/cumulative logit models)
  - We insert also the cluster size n_j among the predictors (Grund et al., 2017)
  - Imputations are carried out iteratively until convergence

Substantive model: modelling teacher beliefs

Teacher beliefs are measured by 20 ordinal items on a 7-point Likert scale. How to model the effect of an ordinal item? Two opposite approaches: (i) the item is treated as continuous (1 parameter), i.e. the effect is assumed to be linear, (ii) the item is treated as nominal (dummy-coding with 6 parameters)

Modelling the effects of teacher beliefs involves two problems:

1. The effect of the items could be non-linear → need a flexible parametrization (but still parsimonious)
2. The items are strongly associated (there is much redundancy) → need to select the most relevant items



After fitting the model with dummy coding (1 dummy for each category), item Q12 does not show a linear effect

Variable selection with regularization

An approach to jointly select the items and the relevant categories of each item is to rely on regularization methods for ordinal predictors (Gertheiss J. and Tutz G., 2010. Sparse modeling of categorical explanatory variables, *The Annals of Applied Statistics*)

In this approach, the ordinal predictors are 'dummy coded' (e.g. 7 categories → 6 dummy variables) and the lasso penalty term is as follows:

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

γ_{kc} corresponds to the parameter of category *c* of predictor *k*

- Categories are merged if some adjacent dummy coefficients are set equal
- A predictor is excluded if all the differences between its coefficients are zero

This approach is equivalent to apply the standard lasso to the transformed dataset where the ordinal predictors are backward-difference coded (split-coding)

$$J(\boldsymbol{\beta}) = \sum_{k=1}^{K} \sum_{c=2}^{C_k} \frac{1}{|\hat{\beta}_{kc}|} |\beta_{kc}|$$

$$\beta_{kc} = \gamma_{kc} - \gamma_{k,c-1}$$

$\hat{\beta}_{kc}$ is an OLS preliminary estimate (adaptive lasso)

Implementation of the algorithm

After split-coding, any package for adaptive lasso can be used.

➢ We exploit the lasso2 command (Stata lassopack procedure)

Ahrens A., Hansen C.B., 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.

✓ The regularization procedure of lasso2 uses the coordinate descent algorithm to minimize the penalized criterion:

$$Q(\boldsymbol{\beta}) = \frac{1}{n} RSS(\boldsymbol{\beta}) + \frac{\lambda}{n} J(\boldsymbol{\beta})$$

✓ The penalty parameter λ is chosen by minimizing the Extended BIC index (Chen, J., Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. Biometrika, 95(3), 759–771)

$$EBIC = n \log(RSS/n) + s \log(n) + 2s \log(p)$$

where *s* is the number of parameters in the fitted model, and *p* is the total number of available predictors.

Combining variable selection and multiple imputation: our strategy

1. For each of the *M* imputed data sets, perform variable selection with adaptive lasso for ordinal predictors
2. Retain the predictors who are selected in at least 50% of the *M* imputed data sets - this threshold works well in the simulation study of Wood A.M., White I.R., Royston P. (2008) How should variable selection be performed with multiply imputed data? *Statistics in Medicine*
3. For each of the *M* imputed data sets, fit the linear random intercept model with the retained predictors
4. Combine the *M* vectors of estimated coefficients and the corresponding standard errors exploiting Rubin rules (Little, R.J.A., Rubin, D.B. (2002) *Statistical Analysis with Missing Data, 2nd Edition,* Wiley)
5. Steps 3 and 4 can be iterated to choose the final model on the basis of statistical tests

Application: selection of PRODID items on teacher practices and beliefs

- The preliminary variable selection step concerns **10 binary** and **20 ordinal** items from the PRODID teacher questionnaire, while the other predictors are included in the model but not penalized.
- Adopting backward-coding for the ordinal items, the total number of parameters for the PRODID items amounts to 130 (6 parameters for each ordinal item + 1 for each binary item).
- Applying the lasso to the *M*=10 imputed data sets with the 50% criterion we retain **5 binary** items and **13 ordinal** items.
- Moreover, for each ordinal item only few parameters were retained, implying collapsing of categories
  - e.g. for item Q15 the lasso retains parameters β_4 and β_6 collapsing the categories as follows: (1,2,3), (4,5), (6,7).

Application: model fit and results

- The random intercept linear model is fitted by maximum likelihood on the $M = 10$ imputed data sets, and the results are combined with Rubin rules.
- The analysis is conducted with Stata (`mixed` and `mi` commands)
- Using the *p*-values derived from Rubin rules, only **1 binary** item and **4 ordinal** items remain significant (*sign of the coefficients in parenthesis*):
  - Teacher practices (binary):
    - Q02 contribution of external experts (**+**)
  - Teacher beliefs (ordinal):
    - Q12 teaching is an exciting experience (**+**)
    - Q15 learning needs cooperation among students (**-**)
    - Q17 student opinions are a key indicator of course quality (**+**)
    - Q27 interest in discuss teaching methods with colleagues (**-**)

Response variable
student rating on teacher ability to involve students

Application: explained level 2 variance

Response variable student rating on teacher ability to involve students

- 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.
  - Fitting the random intercept model without any predictor yields an estimated level 2 variance $\hat{\sigma}_u^2 = 1.3320$
  - After introducing all the predictors except teacher practices and beliefs, it reduces to $\hat{\sigma}_u^2 = 1.2306$ (-8%)
  - The final model gives $\hat{\sigma}_u^2 = 1.0012$, 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.
- While the majority of level 1 variability remains unexplained, i.e. it is most influenced by student unobserved characteristics.

Impact of MI on the estimators variability

To quantify the missing data's influence on the sampling variance of a parameter estimate we can consider the fraction of missing information FMI (Rubin, 1987)

$$\text{FMI} = \left(\frac{V_B + V_B/\text{M}}{V_T} \right)$$

- Level 1 predictors are fully observed and cluster-mean centered, so they are not affected by imputations of level 2 predictors: their FMIs are near zero
- Fully observed level 2 predictors (i.e. teacher and course characteristics) are little affected by imputations, showing FMI between 0.01 and 0.17
- 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%).
- For the imputed covariates with FMI<50% the trade-off between the sampling error increase of a parameter due to MI and its reduction due to data augmentation is favourable, i.e. the relative efficiency is high

Variable selection and MI in multilevel models: open issues

- Variable selection: how to account for clustered observations?
  - Our approach: 1. selection via adaptive lasso for standard linear models; 2. fit a random intercept model using the selected predictors
  - Direct approach: `lmmLasso` of R perform lasso for mixed models (we tried, but we encountered computational problems)
- Imputation step: MICE could be replaced by Joint Modelling (Quartagno and Carpenter, 2016) or by the latent class approach (Vidotto *et al.*, 2018), which give valid inferences for a wider set of model specifications, including non linear effects and interactions.
- How to combine MI and variable selection?
  - Our strategy is simple to face a computationally demanding setting
  - It would be interesting to explore other approaches (a recent review: Zhao and Long, 2017, WIREs Comput. Stat.)

UNIVERSITÀ DEGLI STUDI FIRENZE

Leonardo Grilli
M. Francesca Marino
Omar Paccagnella
Carla Rampichini

UNIVERSITÀ DEGLI STUDI DI PADOVA

Thanks for your attention!

The authors gratefully acknowledge the support of grant SID2016 from the University of Padua *Advances in Multilevel and Longitudinal Modelling* (PI: Omar Paccagnella)

Final workshop of the project: Padua, January 28, 2019
http://amalm.stat.unipd.it/

Draft of the paper https://arxiv.org/abs/1904.05062

References: imputation

- Carpenter, J., & Kenward, M. (2013). *Multiple imputation and its application*. Chichester, United Kingdom: John Wiley & Sons, Ltd.
- Erler NS, Rizopoulos D, van Rosmalen J, Jaddoe VWV, Franco OH and Lesaffre E MEH (2016). Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian*, Statist. Med.*, 35 2955–2974
- Grund S., Ludtke O., Robitzsch A. (2017) 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 Vol. XX, No. X, pp. 1–38
- van Buuren S. (2012). *Flexible Imputation of Missing Data*, Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press Taylor & Francis Group: Boca Raton, FL
- Vermunt, J.K., Van Ginkel J.R., Van der Ark L.A., Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Meth.*, 38, 369-397.
- Vidotto, D., Kaptein, M. C., & Vermunt, J. K. (2015). Multiple imputation of missing categorical data using latent class models: State of art. *Psychological Test and Assessment Modeling*, 57(4), 542-576.

References: variable selection

- Ahrens A., Hansen C.B., 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.
- Gertheiss, J. & Tutz, G. (2009) Penalized regression with ordinal predictors. International Statistical Review, 77, 345-365.
- Gertheiss J. and Tutz G. (2010). Sparse modeling of categorical explanatory variables, The Annals of Applied Statistics, Volume 4, Number 4, 2150-2180.
- Tutz G. and Gertheiss J. (2016). Regularized regression for categorical data, Statistical Modelling, 16(3): 161–200
- Zhao and Long (2017). Variable selection in the presence of missing data: imputation-based methods, WIREs Comput Stat, 9:e1402. doi: 10.1002/wics.1402