Boletín de Estadística e Investigación Operativa
Vol. 34, No. 1, Marzo 2018, pp. 7-24

Estadística

A handful of critical choices in multilevel modelling

Leonardo Grilli and Carla Rampichini
Department of Statistics, Computer Science, Applications
University of Florence
✉ leonardo.grilli@unifi.it, ✉ carla.rampichini@unifi.it

Abstract

Multilevel models are a key tool for the analysis of clustered data in a wide range of fields. The paper discusses a handful of critical choices in multilevel modelling. Some choices are peculiar of the multilevel setting, like the specification of the multilevel structure of the model, cluster-mean centering of the covariates, fixed versus random effects, and the specification of the distribution of the random effects. The paper also considers some choices which are more complicated in the multilevel setting, namely sample size requirements, accounting for the survey design, and handling missing values. Each issue is briefly outlined, referring to the current literature for details and further discussion.

Keywords: hierarchical data, mixed effects models, random effects, variance components.
AMS Subject classifications: 62J05, 62J10.

1. Introduction

Clustered data are often encountered in applied research, e.g. students nested within schools in cross-section data, and occasions nested within subjects in panel or longitudinal data. Multilevel models are the main tool for the analysis of clustered data. Depending on the field, multilevel models are also known as hierarchical, mixed effects, random effects, random coefficients, or variance components. Theoretical and practical aspects of multilevel modelling are presented in several excellent textbooks, including Raudenbush and Bryk (2002), Goldstein (2011), Snijders and Bosker (2012), Rabe-Hasketh and Skrondal (2012), Hox *et al.* (2017). Some alternative methods for clustered data are discussed by McNeish *et al.* (2017).

The paper does not intend to systematically review the wide literature on multilevel models, rather to outline and discuss some critical issues that arise in their specification.

To introduce terms and symbols, let us refer to a two-level hierarchy where level 1 units are indexed by i and level 2 units (clusters) are indexed by j. The structure can be unbalanced, with clusters of different size n_j. In this setting, the random intercept linear model with one covariate at level 1 and one covariate at level 2 is

$$y_{ij} = \alpha + \beta x_{ij} + \gamma z_j + u_j + e_{ij} \tag{1.1}$$

where y_{ij} is the response variable, x_{ij} is the level 1 covariate, z_j is the level 2 covariate, u_j is the level 2 error or *random effect*, and e_{ij} is the level 1 error.

In the base specification, the level 2 errors are independent and identically distributed (thus homoscedastic) across clusters, with normal distribution:

$$u_j \overset{iid}{\sim} N(0, \sigma_u^2) \tag{1.2}$$

The level 1 errors, which are independent of the level 2 errors, are assumed to be independent and identically distributed (thus homoscedastic) across level 1 units, with normal distribution:

$$e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2) \tag{1.3}$$

A further assumption, often not explicitly stated, is the mean independence of the random effect on the covariates (level 2 exogeneity), namely

$$E(u_j \mid x_{1j}, x_{2j}, \ldots, x_{n_j j}, z_j) = 0 \tag{1.4}$$

Exogeneity is needed for unbiased estimation (Ebbes *et al.* 2004; Kim and Frees, 2007; Grilli and Rampichini, 2011).

Multilevel models are a kind of regression models which are intended to account for the correlation structure induced by the arrangement of the units. In any regression model the researcher has to make several choices concerning the explanatory variables, the functional form, the distribution of the errors, and so on. The complex nature of a multilevel model complicates the usual choices and it entails further peculiar choices. The simple random intercept linear model (1.1) can be extended in many ways to accommodate more complex structures, for example by adding further random effects (nested or crossed), by relaxing the level 1 or level 2 homoscedasticity assumptions, by changing the errors distributions. For the sake of simplicity, the discussion of the critical choices in multilevel modelling is exemplified with reference to the *linear* case. However, the main issues are conceptually similar in non-linear multilevel models.

The rest of the paper is organized as follows. Sections 2-5 discuss choices which are peculiar to the multilevel setting: specifying the multilevel structure of the model, cluster-mean centering of the covariates, fixed versus random effects, and specification of the distribution of the random effects. Sections 6-8

consider choices which are more complicated in the multilevel setting: sample size requirements, accounting for the survey design, and handling missing values.

2. Specification of the multilevel structure of the model

There is a wide range of possible multilevel structures. The simplest structure is nested, also known as hierarchical. In a two-level structure level, 1 units are nested into level 2 units (clusters). A key remark is that in a multilevel structure the concept of *unit* is wide: a unit can be a physical entity (an individual, a mice, a product), an organization (a firm, a school), a social institution (a family, a cattle), a geographic area. In addition, a unit can be a measurement, thus panel data models and multivariate models can be interpreted as instances of multilevel models. For example, panel or longitudinal data can be seen as two-level nested data with occasions at level 1 and individuals at level 2; a similar idea applies to multivariate data such as questionnaire surveys, where items are level 1 units and individuals are level 2 units. Panel and multivariate data are peculiar instances of multilevel models since level 1 units are not exchangeable: for example, questionnaire items i and i' are not exchangeable, contrary to students i and i'. In addition, in panel data occasions are ordered in time, a feature that motivates peculiar modelling choices such as autocorrelated residuals (e.g. Skrondal and Rabe-Hesketh, 2008).

Multilevel models can be extended to accommodate nested structures with 3 or more levels by adding random effects at each level. Sometimes the phenomenon under study has a hierarchical structure with many levels, for example, in education it is common to encounter five-level structures (student, class, school, district, geographic area). In principle, one could account for all levels by adding corresponding random effects to the model. However, the inclusion of all the possible levels is not always advisable: indeed, more complex structures give a finer representation of the phenomenon, but they entail problems in estimation, especially for levels with few units (see Section 6). In general, to reduce the number of levels it is preferable to discard higher levels than omit intermediate levels. For example, in a study on peer-effects the class level must be considered, while the school level could be omitted. In general, ignoring top hierarchical levels inflates the variance component at the highest level of the specified model; indeed, this variance component collects all the sources of variance at higher levels (Tranmer and Steel, 2001). To adjust for the correlation between classes of the same school it is possible to use robust standard errors for clustered observations (Rabe-Hesketh and Skrondal, 2006), with the caveat that this correction has a poor performance when the number of schools is small (Cameron and Miller, 2015). When the number of top level units is small, an alternative approach is to use fixed effects at that level, even if this approach precludes the introduction of covariates at top level (see Section

4).

Multilevel models can also deal with non-hierarchical structures. Two important cases are cross-classified and multiple-membership structures.

Cross-classified multilevel models are useful when level 1 units are classified by two non-nested factors (Browne *et al.*, 2001), for example students classified by school and neighbourhood (Leckie, 2009). A simple linear additive cross-classified model is

$$y_{ijk} = \alpha + \beta x_{ijk} + u_j + u_k + e_{ijk} \tag{2.1}$$

where y_{ijk} is the response of student i, belonging to school j and neighbourhood k. In model (2.1) it is assumed that the random effects u_j and u_k are independent with zero means and distinct variances. Cross-classified models are useful also in longitudinal studies where units can change their group membership over time, e.g. repeated measures of student achievement classified by student and school. In such a case, the level 1 unit is the occasion (wave), while the classification factors are the student and the school (Luo and Kwok, 2012).

In multiple membership models each level 1 unit may belong to more than one cluster (e.g. because it moved across the clusters), consequently the random effects enter the model through weights reflecting the hypothesized contribution of each cluster to the outcome of the level 1 unit (Browne *et al.*, 2001). For example, suppose that in a school cycle of 5 years a student spent 4 years in school A and then moved to school B, where she took a final examination for assessing the progress during the whole cycle. It is clearly unfair to ascribe the progress of such a student only to school B, as in a standard multilevel model (Goldstein *et al.*, 2007). Instead, it is reasonable to assume that the progress of such a student is due to school A for 4/5 and to school B for 1/5, though the values of those weights are questionable (Wolff Smith and Beretvas, 2014). Recently, multiple membership models are used to analyse social network data, where individuals may belong to several subgroups (e.g. Tranmer *et al.*, 2014, 2016).

Special considerations are needed in repeated cross-sectional surveys, such as repeated surveys on countries, where a country appears at all surveys, while sampled individuals within the country are different. The model can be specified in many ways, for example: (*i*) a three-level model with individuals nested within survey, and surveys nested within country, or (*ii*) a cross-classified model with individuals cross-classified by country and survey. The implications of different specifications are discussed in Schmidt-Catran and Fairbrother (2016).

To summarize, the structure to be used in the analysis mainly depends on the aim of the research and on the sampling design. For most purposes a simple structure with 2 or 3 levels is appropriate. As usual in statistical modelling, it

is advisable to keep the model as simple as possible: as remarked by DiPrete and Forristal (1994), the imagination of the researchers "can easily outrun the capacity of the data, the computer, and current optimization techniques to provide robust estimates".

3. Cluster mean centering of the covariates

When data have a hierarchical structure, a level 1 covariate can vary both within and between clusters. In general, a covariate x_{ij} can be written as:

$$x_{ij} = (x_{ij} - \overline{x}_j) + \overline{x}_j$$

where $\overline{x}_j = (1/n_j) \sum_{i=1}^{n_j} x_{ij}$ is the cluster mean. The deviation from the cluster mean $(x_{ij} - \overline{x}_j)$ varies only within clusters, whereas the cluster mean $\overline{x}_j$ varies only between clusters. Consequently, the variance of x_{ij} is decomposed into the sum of level 1 and at level 2 variances.

The two components $(x_{ij} - \overline{x}_j)$ and $\overline{x}_j$ may have a different effect on the response y_{ij}. In order to disentangle such effects, these two components must be inserted as distinct covariates into the model. Since the within component is the raw covariate centered with respect to the cluster mean, this kind of model specification is known as *cluster mean centering*.

To sketch the issue, let us consider a two-level linear model with a single continuous covariate x_{ij},

$$y_{ij} = \alpha + \beta_{\text{within}}(x_{ij} - \overline{x}_j) + \beta_{\text{between}}\overline{x}_j + u_j + e_{ij} \tag{3.1}$$

The within and between effects are conceptually different. The within effect β_{within} is the slope in the regression of $(y_{ij} - \overline{y}_j)$ on $(x_{ij} - \overline{x}_j)$, i.e. the effect at the individual level. On the other hand, the between effect β_{between} is the slope in the regression of $\overline{y}_j$ on $\overline{x}_j$, i.e. the effect at the cluster level. The difference $\delta = \beta_{\text{between}} - \beta_{\text{within}}$ is the so called *contextual effect*, which can be directly estimated if model (3.1) is reparametrized as follows:

$$y_{ij} = \alpha + \beta_{\text{within}}x_{ij} + \delta\overline{x}_j + u_j + e_{ij} \tag{3.2}$$

The contextual effect δ summarizes the effect of the context on the individual, which is of central interest in fields such as epidemiology and education.

If the contextual effect δ is not null, but the cluster mean $\overline{x}_j$ is omitted from model (3.2), the model is wrongly specified and the regression coefficient of x_{ij} is not interpretable. This can be seen as a problem of level 2 endogeneity: indeed, in such a case the covariate x_{ij} is correlated with the random effect u_j (Ebbes *et al.*, 2004; Kim and Frees, 2007; Grilli and Rampichini, 2011).

In model (3.2) the estimator of the within-effect β_{within} is unbiased, while

the estimator of the contextual effect δ could still be biased if the cluster mean $\overline{x}_j$ is correlated with the random effect u_j. If the researcher is interested in the estimation of the contextual effect δ, a possible solution is the use of instrumental variables (Castellano *et al.*, 2014). Otherwise, the researcher interested solely in the estimation of the within effect β_{within} can fit model (3.1) or model (3.2), interpreting only β_{within} and ignoring the other regression coefficients. The same estimate of β_{within} can be obtained by fitting a model with only the centered covariate (e.g. Raudenbush and Bryk, 2002):

$$y_{ij} = \alpha + \beta_{\text{within}}(x_{ij} - \overline{x}_j) + u_j + e_{ij} \tag{3.3}$$

Note that models (3.1) and (3.2) give the same estimate of the cluster variance σ^2_u, while model (3.3) yields a higher estimate, since the between-component of the covariate is absorbed by the random effect u_j.

Cluster mean centering is discussed in many textbooks and papers (e.g. Paccagnella, 2006; Enders and Tofighi, 2007). This issue is still source of misunderstandings, e.g. Kelley *et al.* (2017) and the related comment by Bell *et al.* (2017).

4. Fixed versus random effects

The random effect u_j in model (1.1) is a random variable representing unobserved factors acting at level 2 (similarly, e_{ij} represents unobserved factors acting at level 1). Alternatively, unobserved factors acting at level 2 can be treated as unknown fixed quantities, i.e. *fixed effects*: in such a case the random variable u_j is replaced by a set of parameters, say $\alpha_1, \ldots, \alpha_J$, where J is the number of clusters.

The main problem with the random effect approach is the risk of misspecification: indeed, a wrong specification of the conditional distribution of the random effect given the covariates may yield biased inferences. It is therefore crucial to check the assumptions on the random effect and possibly adopt alternative specifications (Snijders and Berkhof, 2008; Drikvandi *et al.*, 2017). For example, the correlation of the random effect with a level 1 covariate can be solved by introducing the corresponding cluster mean (see Section 3). Nonetheless, the analyst can never be sure that the adopted specification is correct.

The fixed effects approach eliminates the mentioned risk of misspecification because there is no need to specify a probability distribution, nor to assume that the effects are uncorrelated with the covariates (exogeneity). A noteworthy feature is that the fixed effects α_j absorb all the between variation: consequently, the covariates can only explain the within variation, thus the regression coefficients are the within effects, regardless of cluster mean centering.

However, the fixed effects approach has several drawbacks. First, it does not allow to include cluster-level covariates: the technical reason is perfect collinearity, while an intuitive explanation is that the fixed effects fully account for the between-cluster variability, so there is no scope for cluster-level explanatory variables. In many fields, the main research question is to find out factors explaining the role of the context on the individual outcome, e.g. how teaching style affects pupil achievement. In this case, the impossibility to include cluster-level covariates is a serious limitation, precluding to answer the research question.

Another drawback of the fixed effects approach is the incidental parameter problem arising in non-linear models, yielding inconsistent estimators of all the parameters (e.g. Wooldridge, 2010). A possible solution is given by conditional maximum likelihood, which is used for example in the conditional logit model. Skrondal and Rabe-Hesketh (2014) discuss the conditional logit model in the framework of panel data, showing that it gives consistent estimators even for some non-ignorable missing data mechanisms.

In some applications, the researcher is interested in predicting cluster-specific effects, e.g. effectiveness of schools or hospitals. In those situations, the random effects model is preferable since it allows to rely on empirical Bayes (shrunken) residuals to make efficient predictions of cluster-specific effects, especially for small clusters (Snijders and Bosker, 2012).

Another point in favour of random effects is that they allow many extensions that are impossible or tricky with the fixed effects approach, such as spatially correlated effects or slopes varying across clusters (Bell and Jones, 2015).

A final consideration in the choice between fixed and random effects pertains to the kind of desired inference. In fact, fixed effects are used for inference on the clusters in the data. On the other hand, random effects allow to make inference on a population of clusters, assuming that the clusters in the data are a random sample from such a population (Snijders and Bosker, 2012). Nonetheless, it may be reasonable to use random effects even if the data include all the clusters of the population (e.g. all the European countries) since random effects yield a more parsimonious description of the observed variability among clusters.

5. Specification of the distribution of the random effects

In a two-level setting, the standard assumptions for the distribution of the random effects are: (*i*) independence across clusters; (*ii*) identical distribution across clusters; (*iii*) normal distribution. If those assumptions are tenable, model estimation and interpretation are straightforward. However, there are situations where one or more of these assumptions are not reasonable, so they should be relaxed.

In some settings, the assumption that the random effects are independent

across clusters is questionable, for example when the clusters are adjacent geographic areas. Indeed, in fields such as disease mapping (Besag *et al.*, 1991) and small area estimation (Rao, 2003) the models have spatially correlated random effects. In the literature on multilevel models, correlated random effects are uncommon. Nonetheless, Browne and Goldstein (2010) considered multilevel models where the higher-level random effects are linked by a suitable correlation structure to be estimated. This is relevant in educational effectiveness, where the performances of nearby schools may be correlated.

The assumption that the random effects have identical distribution across clusters implies constant cluster variance (homoscedasticity), which is too restrictive in some settings. Heteroscedasticity across strata of clusters (e.g. private vs public schools) is handled by stratum-specific random effects (Sani and Grilli, 2011). Heteroscedasticity depending on continuous covariates can be specified by adding random coefficients to level 2 covariates (Snijders and Bosker 2012, Sect. 8.2), or by specifying a linear model for the logarithm of the cluster variance, as in the mixed location scale model of Hedeker *et al.* (2012).

The assumption of normal distribution for the random effects can be overcome in several ways ranging from two extremes: (*i*) a continuous parametric non-normal distribution, and (*ii*) an arbitrary discrete distribution with locations and masses to be estimated. For a review see Grilli and Rampichini (2015).

Random effects with a discrete distribution are useful especially for three purposes: (*i*) checking the shape of the distribution of the random effects and identifying level 2 outliers (i.e. clusters with an extreme value of the random effect); (*ii*) relaxing the parametric assumption on the distribution of the random effects; and (*iii*) classifying the clusters, e.g. classifying the schools according to their effectiveness, or the individuals according to their time patterns.

From a different perspective, a model with random effects having a discrete distribution can be interpreted as a *latent class* multilevel model (Vermunt, 2003), where the clusters are assumed to belong to latent classes with common unobserved components. In this framework, the choice of the number of latent classes is a difficult task (Lukociene *et al.*, 2010). A promising procedure to classify clusters with data-driven selection of the number of classes is represented by Dirichlet process mixtures (Heinzl and Tutz, 2013). An alternative approach to select the number of classes is based on regularization techniques (Tutz and Oelker, 2016).

In the context of repeated measures, discrete random effects or latent classes are the core of Growth Mixture Models (Muthén, 2004; Palardy and Vermunt, 2010) and Latent Markov Models (Bartolucci *et al.*, 2011).

6. Sample size requirements

A multilevel model requires enough level 2 units (clusters) in order to obtain accurate estimates. The minimum depends on the type of model (linear vs non-linear, random intercept vs random slope), on the average size of the clusters and on the true parameter values. Ten clusters may be enough in a favourable situation, like a linear random intercept model with only level 1 covariates estimated on a data set with large clusters. However, the case study at hand is often far from this favourable situation. Note that the number of clusters poses a limitation on the number of cluster-level covariates: for example, a model with 8 cluster-level covariates fitted on a sample with 10 clusters has 2 degrees of freedom at the cluster level, yielding inaccurate estimates of cluster-level parameters.

The sample size requirements are different depending on the target of inference. The less demanding target is to get accurate point estimates of level 1 regression coefficients: in favourable situations 10 clusters of size 2 may be enough. More clusters (say 30 or 50) are needed for accurate estimation of variance components and standard errors. The requirement is higher for models with random slopes. These results are showed e.g. in Bell *et al.* (2012) and the references therein. For non-linear models (e.g. binary responses) more clusters are needed for accurate inference (e.g. Schoeneberger, 2016).

In cross-country surveys the sample size (citizens) is large, but the number of clusters (countries) is small. This case has received attention in the recent literature (Stegmueller, 2013; Bryan and Jenkins, 2015).

To deal with few clusters in the frequentist context, inference for the linear model can be adjusted using restricted maximum likelihood estimation (REML) with a Kenward-Roger correction (McNeish, 2017). Otherwise, a straightforward solution is to specify a fixed effects model, with the limitations highlighted in Section 4.

An alternative approach to handle few clusters is to fit a random effects model with Bayesian methods, which do not rely on asymptotics (Browne and Draper, 2006). The Bayesian approach is powerful, yielding accurate estimates even for a small number of clusters and properly accounting for all the sources of uncertainty. However, it entails computational difficulties and it requires a difficult choice of the priors of the parameters describing the distribution of the random effects (Gelman, 2006; Grilli *et al.* 2015). The comparison of the performance of different approaches dealing with few clusters is the issue of several studies (e.g. Stegmueller, 2013; Elff *et al.*, 2016; McNeish and Stapleton, 2016).

The cluster size is less relevant than the number of clusters (McNeish, 2014). Clusters of size 2 are usually enough for a linear random intercept model; even clusters with a single unit are not an issue, as long as they are not too

many (say more than a half). However, small clusters worsen cluster-specific inferences, for example, the precision of Empirical Bayes predictions of random effects. Moreover, data with small clusters carry limited information on the variance-covariance structure at level 2, so random slopes are likely to be not significant.

The design of multilevel studies requires several choices, depending on the inferential target and the available budget (Snijders and Bosker, 2012, ch. 11). The main objective is to determine the sample size needed at each hierarchical level to have sufficient power and precision for the effects of interest. The optimal allocation of sample units is complicated by the fact that sampling level 1 units within an already selected cluster is usually less expensive than sampling in a new cluster. Special considerations arise in case of multilevel randomized studies, where randomization can be applied at the individual level or at the cluster level (Moerbeek and Teerenstra, 2016).

7. Accounting for the survey design

Multilevel data are often collected through complex survey designs with stratification and multi-stage sampling. The question is whether the model specification or the estimation procedure should explicitly account for the survey design. A first consideration pertains the type of desired inference: *descriptive inference* deals with estimation and testing of descriptive parameters of the surveyed population (e.g. the proportion of pupils who fail a specific test in math), whereas *analytic inference* deals with estimation and testing relationships among variables without reference to the surveyed population – the aim is to generalize to a larger population, which may be hypothetical and is usually rather vaguely defined (e.g. a study on peer effects does not refer to a precisely defined population). Statistical modelling, including multilevel modelling, aims at *analytic inference*, which requires a model-based approach rather than a design-based approach: therefore, the sampling design needs to be taken into account only to the extent it affects model fitting (mainly, point estimates and standard errors).

In a model-based approach the sampling design is a nuisance that can be accommodated in several ways: (*i*) by the model, i.e. the stages of sampling define the hierarchical levels having random effects, and the design variables enter as covariates; (*ii*) by the estimation algorithm, i.e. any unit is weighted by the inverse of the inclusion probability, separately for each hierarchical level; (*iii*) by the standard errors, i.e. robust (sandwich) estimators of the standard errors are exploited to account for clustering and stratification.

Approach (*i*) is in principle straightforward, but a model fully accounting for the sampling design may become unduly complicated. Therefore, it can be convenient to account for the sampling design by the estimation algorithm

(weighting) and/or by robust standard errors. In practice, hybrid approaches are feasible: for example, a stratified three-stage design can be accommodated by a two-level random intercept model fitted by weighted estimation, with robust standard errors accounting for stratification and top-level clustering (Rabe-Hesketh and Skrondal, 2006). The consequences of ignoring stratification and clustering are investigated by Stapleton and Kang (2016).

It is worth to note that weighted estimation in multilevel modelling requires knowledge of the inclusion probabilities at each hierarchical level, e.g. for a two-level model of pupils within schools it is necessary to know the inclusion probability of each school and the inclusion probability of each pupil conditional on the school being sampled.

The use of sampling weights affects the properties of the estimators in two conflicting ways: reduction of the bias and inflation of the variance. Ignoring the sampling weights yield biased estimates only if the survey design is *informative*, namely the inclusion probabilities are related to the model errors. On the other hand, the variance usually increases for most parameters (the increase tends to be large if the weights are highly variable). In practice, it is difficult to decide if weighting is convenient. A preliminary step is the analysis of the sampling design and the data to find clues of possible biases. Then it is recommended to compare weighted and unweighted estimates: a formal test (global or separately for each parameter) can be derived following the principle of the Hausman test, though the test may be unreliable due to the difficulty in the estimation of the covariance matrix of the weighted estimator. Alternatively, weighted and unweighted estimates can be compared with indexes of informativeness (Asparouhov, 2006; Grilli *et al.*, 2016).

A final warning is that weighted estimators work poorly with a low number of clusters at the top level (Primary Sampling Units), say less than 30. In those situations also cluster-robust standard errors are likely to be unsatisfactory (Cameron and Miller, 2015). When clusters have large sizes, like in cross-country research, a solution could be a two-step approach (Achen, 2005) where weighted estimation is performed separately for each cluster.

A thorough treatment of weighting in multilevel modelling, covering both theoretical and technical issues, is in Chapter 14 of Snijders and Bosker (2012).

8. Handling missing values

In applied research missing values are a common issue. The naif approach of using only the observed data (listwise deletion) is not advisable for two main reasons: (*i*) listwise deletion reduces the sample size, and thus the statistical power, and (*ii*) listwise deletion yields biased estimates unless the missing mechanism is MCAR (Missing Completely At Random), namely the probability of a missing value does not depend on the complete data (Seaman et al., 2013).

In most applications it is reasonable to assume that the missing mechanism is MAR (Missing At Random), i.e. conditionally on the observed data the probability of a missing value does not depend on the unobserved data (Seaman et al., 2013). In such a case, the main approaches to deal with missing values are full information maximum likelihood (FIML) and multiple imputation (MI). The FIML approach is theoretically appealing, but it is feasible only in special situations. On the other hand, the MI approach is very flexible and it can be effectively generalized to a wide range of situations.

Multilevel analysis raises special issues. In particular, missing values on the covariates can be both at level 1 and at level 2, and missing values can alter the variance components and the correlations. Multiple imputation has been extended to the multilevel setting to deal with these special issues, following two main approaches: joint modelling and fully conditional specification (FCS), also known as multivariate imputation by chained equations (MICE). See Snijders and Bosker (2012), van Buuren (2012), Carpenter and Kenward (2013), Goldstein *et al.* (2014), Mistler and Enders (2017), Enders *et al.* (2017). The performances of several approaches are compared by Grund *et al.* (2018).

Other methods to deal with missing values in the multilevel framework include MI through latent class models (Vidotto *et al.*, 2015) and the full Bayesian approach (Erler *et al.*, 2016).

References

[1] Achen, C. H. (2005). Two-step hierarchical estimation: Beyond regression analysis. *Political Analysis*, **13**, 447–456.

[2] Asparouhov, T. (2006). General Multi-Level Modeling with Sampling Weights. *Communications in Statistics - Theory and Methods*, **35**, 439–460.

[3] Bartolucci, F., Pennoni, F., Vittadini, G. (2011). Assessment of School Performance Through a Multilevel Latent Markov Rasch Model. *Journal of Educational and Behavioral Statistics*, **36**, 491–522.

[4] Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., Ferron, J. M. (2012). How low can you go? An investigation of the influence of sample size and model complexity on point and interval estimates in two-level linear models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, **10**, 1-11.

[5] Bell, A., Jones, K., Fairbrother, M. (2017). Understanding and misunderstanding group mean centering: a commentary on Kelley et al.'s dangerous practice, to appear in *Quality and Quantity*, https://doi.org/10.1007/s11135-017-0593-5.

[6] Bell, A., Jones, K. (2015). Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods*, **3**, 133–153.

[7] Besag, J., York, J., Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.

[8] Browne, W.J., Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models, *Bayesian Analysis*, **1**, 673–514.

[9] Browne, W., Goldstein, H., Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models, *Statistical Modelling*, **1**, 103–124.

[10] Browne, W., Goldstein, H. (2010). MCMC Sampling for a Multilevel Model With Nonindependent Residuals Within and Between Cluster Units. *Journal of Educational and Behavioral Statistics*, **35**, 453–473.

[11] Bryan, M.L., Jenkins, S.P. (2015). Multilevel Modelling of Country Effects: A Cautionary Tale. *European Sociological Review*, **32**, 3–22.

[12] Cameron, A. C., Miller, D.L. (2015). A Practitioner's Guide to Cluster-Robust Inference, *Journal of Human Resources*, **31**, 317–372.

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

[14] Castellano, K., Rabe-Hesketh, S., Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. *Journal of Educational and Behavioral Statistics*, **39**, 333–367.

[15] DiPrete, T.A., Forristal, J.D. (1994). Multilevel Models: Methods and Substance, *Annual Review of Sociology*, **20**, 331–357.

[16] Drikvandi, R., Verbeke, G., Molenberghs, G. (2017). Diagnosing Misspecification of the Random-Effects Distribution in Mixed Models. *Biometrika*, **73**, 63–71.

[17] Ebbes, P., Bockenholt, U., Wedel, M. (2004). Regressor and random-effects dependencies in multilevel models. *Statistica Neerlandica*, **58**, 161–178.

[18] Elff, M., Heisig, J.P., Schaeffer, M., Shikano, S. (2016). No Need to Turn Bayesian in Multilevel Analysis with Few Clusters: How Frequentist Methods Provide Unbiased Estimates and Accurate Inference. *SocArXiv Preprint*, Version 2. URL: https://osf.io/preprints/socarxiv/z65s4

[19] Enders, C.K., Tofighi, D. (2007). Centering Predictor Variables in Cross-Sectional Multilevel Models: A New Look at an Old Issue. *Psychological Methods*, **12**, 121–138.

[20] Enders, C.K., Keller, B.T., Levy, R. (2017) A Fully Conditional Specification Approach to Multilevel Imputation of Categorical and Continuous Variables. *Psychological Methods*, http://dx.doi.org/10.1037/met0000148.

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

[22] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). *Bayesian Analysis*, **1**, 515–534.

[23] Goldstein, H. (2011). *Multilevel Statistical Models. 4th edition.* John Wiley & Sons.

[24] Goldstein, H., Burgess, S., McConnell, B. (2007). Modelling the effect of pupil mobility on school differences in educational achievement, *Journal of the Royal Statistical Society Series A*, **170**, 941–954.

[25] Goldstein, H., Carpenter, J. R., 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*, **177**, 553–564.

[26] Grilli, L., Metelli, S., Rampichini, C. (2015). Bayesian estimation with INLA for logistic multilevel models. *Journal of Statistical Computation and Simulation*, **85**, 2718–2726.

[27] Grilli, L., Rampichini, C. (2009). Multilevel models for the evaluation of educational institutions: a review. In: Monari P., Bini M., Piccolo D., Salmaso L. (eds.) *Statistical Methods for the Evaluation of Educational Services and Quality of Products*, pp. 61–80. Physica-Verlag.

[28] Grilli, L., Rampichini, C. (2011). The role of sample cluster means in multilevel models: a view on endogeneity and measurement error issues. *Methodology*, **7**, 121–133.

[29] Grilli, L., Rampichini, C. (2015). Specification of random effects in multilevel models: a review. *Quality & Quantity*, **49**, 967–976.

[30] Grilli, L., Pennoni, F., Rampichini, C., Romeo, I. (2016). Exploiting TIMSS and PIRLS combined data: multivariate multilevel modelling of student achievement. *The Annals of Applied Statistics*, **10**, 2405–2426.

[31] Grund, S., Ludtke, O., Robitzsch, A. (2018). Multiple Imputation of Missing Data for Multilevel Models Simulations and Recommendations, *Organizational Research Methods*, **21**, 111–149.

[32] Hedeker, D., Mermelstein, R.J., Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine*, **31**, 3328–3336.

[33] Heinzl, F., Tutz, G. (2013). Clustering in linear mixed models with approximate Dirichlet process mixtures using EM algorithm. *Statistical Modelling*, **13**, 41–67.

[34] Hox, J.J., Moerbeek, M., van de Schoot, R. (2017). *Multilevel Analysis: Techniques and Applications, Third Edition*. Routledge.

[35] Kelley, J., Evans, M.D.R., Lowman, J., Lykes, V. (2017). Group-mean-centering independent variables in multilevel models is dangerous. *Quality & Quantity*, **51**, 261–283.

[36] Kim, J.S., Frees, E.W. (2007). Multilevel Modeling with Correlated Effects. *Psychometrika*, **72**, 505–533.

[37] Leckie, G. (2009). The complexity of school and neighbourhood effects and movements of pupils on school differences in models of educational achievement, *Journal of the Royal Statistical Society Series A*, **172**, 537–554.

[38] Seaman, S., Galati, J., Jackson, D., Carlin, J. (2013). What Is Meant by "Missing at Random"? *Statistical Science*, **28**, 257–268.

[39] Lukociene, O., Varriale, R., Vermunt, J.K. (2010). The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociological Methodology*, **40**, 247–283.

[40] Luo, W., Kwok, O. (2012). The Consequences of Ignoring Individuals' Mobility in Multilevel Growth Models: A Monte Carlo Study. *Journal of Educational and Behavioral Statistics*, **37**, 31–56.

[41] McNeish, D. M. (2014). Modeling Sparsely Clustered Data: Design-Based, Model-Based, and Single-Level Methods. *Psychological Methods*, **19**, 552–563.

[42] McNeish, D. M. (2017). Small Sample Methods for Multilevel Modeling: A Colloquial Elucidation of REML and the Kenward-Roger Correction, *Multivariate Behavioral Research*, **52**, 661–670.

[43] McNeish, D.M., Stapleton, L.M. (2016). Modeling Clustered Data with Very Few Clusters, *Multivariate Behavioral Research*, **51**, 495–518.

[44] McNeish, D., Stapleton, L. M., Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, **22**, 114–140.

[45] Mistler, S. A., Enders, C. K. (2016). A comparison of joint model and fully conditional specification imputation for multilevel missing data. *Journal of Educational and Behavioral Statistics*, **42**, 432–466.

[46] Moerbeek, M., Teerenstra, S. (2016). *Power Analysis of Trials with Multilevel Data*. CRC Press.

[47] Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In: Kaplan D. (ed.) *Handbook of quantitative methodology for the social sciences*, 345–368, Sage.

[48] Paccagnella, O. (2006). Centering or not centering in multilevel models. The Role of the Group Mean and the Assessment of Group Effects. *Evaluation Review*, **30**, 66–85.

[49] Palardy, G., Vermunt, J.K. (2010). Multilevel growth mixture models for classifying groups. *Journal of Educational and Behavioral Statistics*, **35**, 532–565.

[50] Rabe-Hesketh, S., Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society Series A*, **169**, 805–827.

[51] Rabe-Hesketh, S., Skrondal, A. (2012). *Multilevel and Longitudinal Modeling Using Stata, Third Edition*. Stata Press.

[52] Rao, J.N.K. (2003). *Small area estimation*. LinkHoboken (N.J.): John Wiley & sons.

[53] Raudenbush, S.W., Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage.

[54] Sani, C., Grilli, L. (2011). Differential variability of test scores among schools: a multilevel analysis of the fifth-grade Invalsi test using heteroscedastic random effects. *Journal of Applied Quantitative Methods*, **6**, 88–99.

[55] Schmidt-Catran, A.W., Fairbrother, M. (2016). The Random Effects in Multilevel Models: Getting Them Wrong and Getting Them Right. *European Sociological Review*, **32**, 23–38.

[56] Schoeneberger, J. A. (2016). The impact of sample size and other factors when estimating multilevel logistic models. *Journal of Experimental Education*, **84**, 373–397.

[57] Snijders, T.A.B., Berkhof, J. (2008). Diagnostic checks for multilevel models, in Jan de Leeuw and Erik Meijer (Eds.), *Handbook of Multilevel Analysis*, Springer, New York.

[58] Snijders, T.A.B., Bosker, R.J. (2012). *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modelling. 2nd edition.* Sage.

[59] Skrondal, A., Rabe-Hesketh, S. (2008). Multilevel and Related Models for Longitudinal Data, in J. de Leeuw, E. Meijer (eds.), *Handbook of Multilevel Analysis*, Springer.

[60] Skrondal, A., Rabe-Hesketh, S. (2014). Protective estimation of mixed-effects logistic regression when data are not missing at random. *Biometrika*, **101**, 175–188.

[61] Stapleton, L.M., Kang, Y. (2016). Design Effects of Multilevel Estimates From National Probability Samples. *Sociological Methods & Research*, 1–28.

[62] Stegmueller, D. (2013). How Many Countries for Multilevel Modeling? A Comparison of Frequentist and Bayesian Approaches. *American Journal of Political Science*, **57**, 748–761.

[63] Tranmer, M., Steel, D. G. (2001). Ignoring a Level in a Multilevel Model: Evidence from UK Census Data, *Environment and Planning A*, **33**, 941–948.

[64] Tranmer, M., Steel, D., Browne, W.J. (2014). Multiple-membership multiple-classification models for social network and group dependences. *Journal of the Royal Statistical Society Series A*, **177**, 439–455.

[65] Tranmer, M., Pallotti, F., Lomi, A. (2016). The embeddedness of organizational performance: Multiple Membership Multiple Classification Models for the analysis of multilevel networks. *Social Networks*, **44**, 269–280.

[66] Tutz, G., Oelker, M.-R. (2016). Modelling Clustered Heterogeneity: Fixed Effects, Random Effects and Mixtures. *International Statistical Review*, **85**, 204–227.

[67] van Buuren, S. (2012). *Flexible Imputation of Missing Data.* Chapman & Hall/CRC.

[68] Vermunt, J.K. (2003). Multilevel latent class models. *Sociological methodology*, **33**, 213–239.

[69] 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**, 542–576.

[70] Wolff Smith, L. J., Beretvas, S. N. (2014). The impact of using incorrect weights with the multiple membership random effects model. *Methodology*, **10**, 31–42.

[71] Wooldridge, J.M., (2010). *Econometric Analysis of Cross Section and Panel Data, second edition*. MIT Press.