Longitudinal Modeling of a material deprivation score with missing data in a rotating panel

In recent years the use and the modeling of longitudinal data have been quite common in various areas of knowledge, although the existence of missing data has been outlined in the literature as one of the most frequent problems to be addressed.

The omissions that can be found consist of different types: i) item non response; ii) unit non response; or iii) attrition. The omissions may also result from the design of the study, as is the case of a planned missing design. In a planned missing design, the missing data occur as on option of the researcher in order to increase the quality of the data available (Graham et al, 2006).

The survey on living conditions and household income (ICOR) is an annual panel that was implemented with the aim of ensuring the Portuguese participation in the European database called EU-SILC (European Statistics on Income and Living Conditions). The ICOR is a rotating panel with a dynamic rotation of ¼ of the sample in each year. Thus, no household or individual will remain in the sample for more than four consecutive years. In this sense, this type of panel configures a case of missing data by design, since a fraction of the data is removed from the sample every year. However ICOR contains other types of missing data, namely, households leaving the study before the fourth year, households without information at the selection moment or even households with intermittent nonresponse.

According to Rubin (1976) there are three missing data mechanisms: 1) missing completely at random (MCAR), if the probability of missing data on a variable is unrelated to other measured variables and to the would-be values of that variable; 2) missing at random (MAR), when the probability of missing data on a variable is related to some other measured variable in the analysis model but not to the would-be values of that variable; and 3) missing not at random (MNAR), if the probability of missing data depends on the would-be value of that variable.

Due to the design adopted, part of the ICOR missing data mechanism can be assumed as missing at random. However other omissions exist, for which it is not possible to assume such a mechanism.

Traditionally, research on material deprivation published in the specialized literature, such as Nolan and Whelan (2010) and Guio (2009), tends to use a statistical cross-sectional approach

to analyzing and modeling this concept. However, existing longitudinal databases offer an opportunity to understand the complexity of the phenomenon and how it evolves over time.

Latent growth curve models are a statistical technique frequently used to study change over time, using longitudinal data. By estimating a latent trajectory over time, this type of modeling allows the study of change both individually or at the household level (Bollen and Curran, 2006). The main parameters of interest in these models are the means and the variance of the random effects.

With ICOR, using a latent growth curve model in order to explain growth trajectories of households' material deprivation over four time points is only possible for 25% of the sample. In order to increase the percentage of households in the study it is required to handle missing data resulting from ICOR design.

When a MAR or a MCAR missing data mechanism can be assumed, the full information maximum likelihood estimation method is one of the state-of-the-art missing data techniques for model estimation (Schaffer & Graham, 2002), as it yields accurate estimates for model parameters.

Full information maximum likelihood method uses all the available data to obtain the parameter estimates. This approach estimates a likelihood function for each individual based on the variables that are present, meaning that all available data is considered (Enders, 2010). However in the presence of a MNAR mechanism, the method produces biased estimates and other approaches should be considered, in particularly, selection models and pattern mixture models.

Selection models for longitudinal data combine a growth curve model with a set of regression equations that predict missingness. According to Wu and Carroll's (1988) model, the repeated measures are indirectly related to the response probabilities through the random effects of the latent growth curve model. In contrast, Diggle and Kenward's (1994) selection model approach directly relates the probability of missing data to the repeated measures, at a particular wave. These two models require different assumptions and may produce different estimates.

The identification of this type of models is done by distributional assumptions. In Wu and Carroll's model approach, distributional assumptions are established for the random effects, whereas in Diggle and Kenward's approach these distributional assumptions are made for the

repeated measures. Additionally, Wu and Carroll's model assumes that the repeated measures and the missing data indicators are conditionally independent, given the random effects.

Similarly to selection models, the pattern mixture models approach integrates a model for the missing data into the analysis, but following a different approach. The pattern mixture analysis stratifies the sample into subgroups sharing the same missing data pattern and estimates a growth model separately within each pattern. Finally, a weighted average of the obtained estimated values for each subgroup is computed, in order to estimate the population growth trajectory.

The pattern mixture model is often not identified and therefore one or more parameters cannot be estimated. In fact, the model is identified only for the subgroup of participants with complete data. In order to ensure identification, restrictions for the inestimable parameters in the model are required; and different restrictions may yield different results.

The aim of this study is to model a material deprivation indicator over time using latent growth curve models, allowing for different missing data patterns present in ICOR. Different approaches of dealing with missing data are used: i) full information maximum likelihood; ii) selection models; and iii) pattern mixture models. Portuguese data from the ICOR, for the period from 2006 to 2009, are considered and the statistical modeling is conducted using the statistical software Mplus 6 (Muthén and Muthén, 1998-2010).