ELSEVIER

CrossMark

Journal of Clinical Epidemiology 68 (2015) 637–645

Journal of Clinical Epidemiology

Including auxiliary item information in longitudinal data analyses improved handling missing questionnaire outcome data

Iris Eekhout[a,b,c,*], Craig K. Enders[d], Jos W.R. Twisk[a,b,c], Michiel R. de Boer[c], Henrica C.W. de Vet[a,b], Martijn W. Heymans[a,b,c]

[a]*Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, De Boelelaan 1089a, 1081 HV, The Netherlands*
[b]*EMGO Institute for Health and Care Research, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands*
[c]*Department of Methodology and Applied Biostatistics, Faculty of Earth and Life Sciences, Institute for Health Sciences, VU University, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands*
[d]*Department of Psychology, Arizona State University, Box 871104 Tempe AZ 85287-1104, USA*

Accepted 21 January 2015; Published online 28 January 2015

Abstract

Objectives: Previous studies show that missing values in multi-item questionnaires can best be handled at item score level. The aim of this study was to demonstrate two novel methods for dealing with incomplete item scores in outcome variables in longitudinal studies. The performance of these methods was previously examined in a simulation study. The two methods incorporate item information at the background when simultaneously the study outcomes are estimated.

Study Design and Setting: The investigated methods include the item scores or a summary of a parcel of available item scores as auxiliary variables while using the total score of the multi-item questionnaire as the main focus of the analysis in a latent growth model. That way the items help estimating the incomplete information of the total scores. The methods are demonstrated in two empirical data sets.

Results: Including the item information results in more precise outcomes in terms of regression coefficient estimates and standard errors, compared with not including item information in the analysis.

Conclusion: The inclusion of a parcel summary is an efficient method that does not overcomplicate longitudinal growth estimates. Therefore, it is recommended in situations where multi-item questionnaires are used as outcome measure in longitudinal clinical studies with incomplete scores because of missing item scores.

Keywords: Missing data; Longitudinal data; Multi-item questionnaire; Auxiliary variables; Full information maximum likelihood; Methods; Latent growth modeling; Structural equation modeling

1. Introduction

Many medical and epidemiologic longitudinal studies use patient-reported outcomes such as quality of life as the main focus of their analyses. These patient-reported outcomes are often repeatedly measured by a multi-item questionnaire. The item scores of the questionnaire are summed or averaged to a total score to represent the outcome of interest. In case respondents do not fill out all the questions in a multi-item questionnaire, the calculation of the total scores is impaired. As a solution, manuals of multi-item questionnaires often advise to average over the available items (e.g., Refs. [1,2]), otherwise known as person mean imputation. Averaging over the available items is algebraically identical to substituting a person's mean item response. This solution can result in biased analysis results, especially when data are not missing completely at random (MCAR) [3,4]. Another option for handling missing data values is to apply a complete-case analysis. In that method, only respondents who have all item scores observed are included in the analysis. This method only results in unbiased analyses when data are MCAR. A complete-case analysis always results in a decreased sample size, so power will be suboptimal in all situations. Nevertheless, this method is most often applied in epidemiologic studies [5].

Conflict of interest: None.

Funding: This work was financially supported by EMGO Institute of Health and Care Research.

* Corresponding author. Department Epidemiology and Biostatistics, VU University Medical Center (F-vleugel), De Boelelaan 1089a, 1081 HV Amsterdam, The Netherlands. Tel.: +31204446040; fax: +31204444645.

E-mail address: i.eekhout@vumc.nl (I. Eekhout).

http://dx.doi.org/10.1016/j.jclinepi.2015.01.012
0895-4356/

What is new?

Key findings

- Including the item information as auxiliary variables in a latent growth model with missing data in the outcome which is assessed with a multi-item questionnaire increases the power and precision of the growth parameters.
- Including item information as parcel summary scores and not as separate items in the latent growth model largely simplifies model estimation without sacrificing accuracy, power and precision.

What this adds to what is known?

- Estimating models by full information maximum likelihood is an advanced method to handle missing data, which produces unbiased regression estimates in MAR data in the outcome of a latent growth model. When outcomes are missing due to missing item scores, including the item information in these models improves the precision of growth estimates. This study shows that this works in empirical data situations and demonstrates how this can improve study conclusions.
- This paper explains how item information can be included in the auxiliary part of a latent growth model.

What is the implication, what should change now?

- When total scores are incomplete due to missing item scores in multi-item questionnaires in a longitudinal growth analysis, item information should be incorporated in the model estimation by using the item information as a parcel summary score in order to get the most accurate and precise estimates.

More advanced methods to handle missing data are multiple imputation or full information maximum likelihood (FIML). Both methods use all observed data in the analyses. In multiple imputation, the missing values are replaced by imputed values. A regression model estimates predicted scores for the incomplete values and random error, drawn from a normal distribution around the estimated value, is added to the predicted score to account for uncertainty around the imputed values. This imputation process is repeated multiple times resulting in multiple imputed data sets. Subsequently, the data analysis is performed on each of these imputed data sets. The multiple results from these data sets are pooled into one final analysis result [6–8]. In FIML, missing values are not replaced or imputed; instead, all available data are used to estimate the population parameters with the highest likelihood of producing the sample data. Both multiple imputation and FIML perform well when the probability of missing data is related to other variables in the data, which is known as missing at random (MAR) [9]. Furthermore, with these techniques, model estimations are generally unbiased and without loss of power.

In a multi-item questionnaire, total scores may be missing because of missing item scores. In that case, there are two main approaches to handle the missing data. Missing data can be handled at the item level or at the total score level of the multi-item questionnaire. The missings are handled at the item level when a missing data method is applied to the incomplete item scores first and then the total scores are calculated (e.g., by summing completed item responses) and used for the analysis. Handling the missings at the total score level means that the total scores will be incomplete when one or more item scores are missing. The missing data handling method is applied to these total scores directly. Previous studies have shown that it is most beneficial to handle the missing data in a multi-item questionnaire at the item level. Handling missing item scores at the item level improves precision [3,4]. In the context of multiple imputation, it is quite straightforward to handle the missings at the item level. The item scores are imputed in the imputation model, and after the imputation part, the item scores are summed to the total scores in each of the imputed data sets, which are used for the analysis. However, when the number of items is very large, for example in longitudinal studies where item scores from multiple time points are included in the analysis, multiple imputation of the item scores might cause complications. When the number of items in the study gets close to the sample size, there is not enough information in the data to estimate the imputation model parameters. For example, in a study where a multi-item questionnaire with 20 items is measured at six time points, the total number of variables in an imputation model would be at least 120. Green [10] described a rule of thumb where the sample size should be larger than $53 + k$ to do a regression analysis for a medium effect size (i.e., 0.13), where k is the number of predictors. In the example, we outline below with 120 variables, the minimum sample size should then be 173. Hence, the number of variables in an imputation model could easily exceed the maximum allowed number in a longitudinal study with many time points and a multi-item questionnaire as outcome measure. Moreover, when outcomes are measured at multiple time points in a longitudinal study, it might be feasible to analyze the data with a longitudinal analysis method such as a latent growth model. Usually, these models are estimated with FIML, which produces unbiased model estimates when missing outcomes are MAR. If the missing data are only in the outcome, handling the missing data by multiple imputation or by FIML in a longitudinal model will yield similar results when the variables in the imputation model are the same as the variables in the longitudinal model [11,12].