Journal of Clinical Epidemiology 102 (2018) 107−114

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Guidelines for multiple imputations in repeated measurements with time-dependent covariates: a case study

Frans E.S. Tan[a,*,1], Shahab Jolani[a,1], Hilde Verbeek[b]

[a]Department of Methodology and Statistics, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, The Netherlands
[b]Department of Health Service Research, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, The Netherlands

Accepted 14 June 2018; Published online 28 June 2018

Abstract

Objectives: We provide guidelines for handling the most common missing data problems in repeated measurements in observational studies and deal with practicalities in producing imputations when there are many partly missing time-varying variables and repeated measurements.

Study Design and Setting: The Maastricht Study on long-term dementia care environments was used as a case study. The data contain 84 momentary assessments for each of 115 participants. A continuous outcome and several time-varying covariates were involved containing missing observations varying from 4% to 25% per time point. A multiple imputation procedure is advocated with restrictions imposed on the relation within and between partially missing variables over time.

Results: Multiple imputation is a better approach to deal with missing observations in both outcome and independent variables. Furthermore, using the statistical package R-MICE, it is possible to deal with the limitations of current statistical software in imputation of missing observations in more complex data.

Conclusion: In observational studies, the direct likelihood approach (i.e., the standard longitudinal data methods) is sufficient to obtain valid inferences in the presence of missing data only in the outcome. In contrast, multiple imputation is required when dealing with partly missing time-varying covariates and repeated measurements.

Keywords: Longitudinal design; Multiple imputation; Observational study; Partly missing time-varying covariates; Overparametrization; R-MICE

1. Introduction

A major advantage of analyzing longitudinal data over cross-sectional data is the possibility to describe individual profiles over time. Because characteristics of subjects may vary over time, measuring the outcome and time-varying characteristics of the subjects repeatedly enables us to better evaluate the effect of them on the outcome for an arbitrary subject [1]. There are many examples, for instance in health care practice, that demonstrate the merits of longitudinal data [2−4]. However, analyzing longitudinal data typically needs advanced approaches when compared to standard cross-sectional data.

Missing data are one of the central problems that one encounters during the analysis of longitudinal data. Subjects may drop out due to, for example, sudden severe illness, death, or inability to locate by the researchers, or a measurement may be missing due to reasons that are unknown to or known but not measured by the researcher. Missing data are a unique challenge all researchers face from time to time, especially those in health care practice [5]. As research designs have become more complex and often multicentered, the problem of missing data has become much more common and complicated. Therefore, statisticians have been addressing this problem over decades and developed solutions that can stand the scrutiny of statistical theory [6−8].

Popular solutions include excluding from the analysis those subjects who have missing observations (i.e., complete cases analysis), simple substitution methods, and advanced approaches like the direct maximum likelihood and multiple imputation (MI) [9]. Although applied researchers may know the existence of these methods, they

Conflict of interest: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

[1] The first two authors contributed equally.

* Corresponding author: Frans E.S. Tan PhD, Associate Professor Methodology and Statistics, Maastricht University, Faculty of Health, Medicine and Life Sciences, CAPHRI Care and Public Health Research Institute, Department of Methodology and Statistics, P.O. Box 616, 6200 MD, Maastricht, The Netherlands. Tel.: +31433882278; Fax: +31433618388.

E-mail address: frans.tan@maastrichtuniversity.nl (F.E.S. Tan).

https://doi.org/10.1016/j.jclinepi.2018.06.006
0895-4356/

What is new?

Key findings

- When analyzing longitudinal data with missing observations, two situations require different approaches. If the missing data are in the outcome only (and the independent variables are fully observed), the direct likelihood method will produce unbiased estimates under the missing at random assumption, and thus multiple imputation is not necessary. If, on the other hand, some of the independent variables contain missing observations too, imputation of missing data is then advantageous.
- A problem arises if there are more columns (variables per time point) than rows (subjects) when the data are constructed for the imputation purpose (i.e. the data are converted to the wide format). With no restrictions, imputing missing data cannot be performed and any software packages will simply crash or stop imputing. Therefore extra restrictions should be imposed while preserving as much as possible the correlation structure of the data, given the imputation model.
- The R- MICE package is useful to successfully deal with such complex longitudinal data.

What this adds to what is known?

- Analysis of the aforementioned complex observational longitudinal data, with many repeated measurements and partly missing time-varying covariates, can be analyzed using the R-MICE package by imposing extra restrictions on the relation within partly missing variables over time.
- When missing data are in the independent variables, the direct likelihood removes subjects with missing observations, which results in biased estimates.

What is the implication and what should change now?

- Care should be taken when analyzing longitudinal data with partly missing observations in the covariates. Moreover, standard software like SPSS and SAS may fail to deliver estimates if there are many time points and time- varying covariates. The guidelines as proposed in the article may be useful for a successful analysis.

may be less aware of the advantages and disadvantages of them depending on the design and underlying missing data mechanisms. Moreover, longitudinal data may have many time points and often contain time-varying independent variables with missing observations [10] so that imputation of missing data using standard software like SPSS and SAS may fail in such complex designs.

The purpose of this article is to provide researchers with practical guidelines to handle the most common missing repeated measurements data problems in observational studies. Many researchers, for example, in health care research and health services, use standard techniques as offered in software like SPSS without realizing the problems that may occur in their particular data. We specifically aim to address

- The important problem of how to analyze longitudinal data if there are missing observations in the outcome only and/or if missing observations are extended to independent variables too. These two situations require different approaches.
- Practicalities in producing imputations when there are many time-varying variables and repeated measurements, such that the imputation task will be impossible without making extra restrictions.
- The difficulties with common and ready-to-use imputation routines in statistical packages SPSS, SAS, and R.

In Section 2, we introduce the Maastricht Study on long-term dementia care environments (MLTD) as a case study and elaborate on its missing data structure. Using this structure as a reference, several potential problems have been considered. In Section 3, a brief review of possible solutions to handle missing data is given. Moreover, a limited simulation study is conducted to further elaborate on performance of different methods based on bias and coverage aspects of the estimates. In Section 4, an outline is given about the statistical analysis of the MLTD study. In Subsection 4.1, tips and tricks are given of how to implement the state-of-the-art method to handle missing observations. In Subsection 4.2, we describe the software limitations by comparing SPSS, SAS, and R-MICE. In Section 4.3 the suggested approach to deal with missing observations is applied to the MLTD study and the results are presented. In section 5, the article ends with a discussion.

2. Missing data structure of the MLTD study

As a motivation example, the MLTD study has a longitudinal design aiming at investigating the effect of innovative dementia care environments (i.e., small scale, homelike) in comparison with traditional nursing homes (large scale) on residents' daily life [3]. In this case study, we are interested to compare the mood between the elderly living in traditional large-scale wards (LSW = 1; 29 wards) and innovative small-scale wards (LSW = 0; 86 wards). A randomized observation schedule was performed, such that