Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 52 (2008) 1533–1548

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data

Caroline Beunckens[a,*], Cristina Sotto[a,b], Geert Molenberghs[a]

[a] *Center for Statistics, Hasselt University, Agoralaan 1, Building D, 3590 Diepenbeek, Belgium*
[b] *School of Statistics, University of the Philippines, Diliman, Quezon City, Philippines*

Received 8 March 2006; received in revised form 17 April 2007; accepted 30 April 2007
Available online 18 May 2007

Abstract

Missingness frequently complicates the analysis of longitudinal data. A popular solution for dealing with incomplete longitudinal data is the use of likelihood-based methods, when, for example, linear, generalized linear, or non-linear mixed models are considered, due to their validity under the assumption of *missing at random* (MAR). Semi-parametric methods such as generalized estimating equations (GEEs) offer another attractive approach but require the assumption of *missing completely at random* (MCAR). Weighted GEE (WGEE) has been proposed as an elegant way to ensure validity under MAR. Alternatively, multiple imputation (MI) can be used to pre-process incomplete data, after which GEE is applied (MI-GEE). Focusing on incomplete binary repeated measures, both methods are compared using the so-called asymptotic, as well as small-sample, simulations, in a variety of correctly specified as well as incorrectly specified models. In spite of the asymptotic unbiasedness of WGEE, results provide striking evidence that MI-GEE is both less biased and more accurate in the small to moderate sample sizes which typically arise in clinical trials.

Keywords: Missing at random; Weighted GEE; Multiple imputation GEE; Asymptotic bias

1. Introduction

Longitudinal binary, or in general non-Gaussian, data are common in biomedical research and beyond. A typical study, for instance, would consist of repeatedly observing the presence or absence of some characteristic, taken in relation to covariates of interest. Data arising from such investigations, however, are often prone to incompleteness, or missingness. In the context of longitudinal studies, missingness predominantly occurs in the form of dropout, in which subjects fail to complete the study for one reason or another. The focus of this paper will be on this type of missingness. In what follows, we will discuss methodology that applies to all non-Gaussian settings, but illustrations and simulations will be confined to the prevalent binary case.

The nature of the dropout mechanism affects both the analysis and interpretation of the remaining data. Since one can almost never be certain about the cause of dropout, certain assumptions have to be made. Therefore, when referring to the missingness process, we will use the terminology introduced by Rubin (1976) and Little and Rubin (1987).

[*] Corresponding author. Tel.: +32 11268257; fax: +32 11268299.
E-mail address: caroline.beunckens@uhasselt.be (C. Beunckens).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.04.020

A non-response process is said to be *missing completely at random* (MCAR) if the missingness is independent of both unobserved and observed data, and *missing at random* (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements. A process that is neither MCAR nor MAR is termed *non-random* (MNAR). Note that specific names for these mechanisms for the case of longitudinal data were cornered by Diggle and Kenward (1994). Moreover, Little (1995) further splits the MCAR case in situations where missingness is independent of both outcomes and covariates on the one hand, and cases where missingness is covariate-dependent only. For reasons of simplicity and generality, we prefer to retain the generic MCAR–MAR–MNAR terminology. Full details can be found in Molenberghs and Kenward (2007). In the context of likelihood inference, and when the parameters describing the measurement process are functionally independent of the parameters describing the missingness process, MCAR and MAR are *ignorable*, while an MNAR process is non-ignorable. This is not the case for frequentist inference, where the stronger condition of MCAR is required to ensure ignorability (Rubin, 1976). Indeed, frequentist methods, such as standard generalized estimating equations (GEEs), for which dropout does not need to be modelled, are only valid under the restrictive MCAR assumption. *Weighted generalized estimating equations* (WGEEs) and *multiple imputation based generalized estimating equations* (MI-GEEs) are two possible alternatives that make it possible to model the data under the MAR missingness mechanism. However, in both methods, dropout needs to be addressed, either by means of a dropout model for WGEE or by an imputation model for MI-GEE, meaning the missing-data mechanism is then not ignorable.

A general taxonomy of models for longitudinal non-Gaussian data consists of three families: marginal, random-effects, and conditional models. Within these model families, a broad set of methods are available, although the marginal and random-effect models are most often used in longitudinal non-Gaussian settings. Such random-effect models, known as generalized linear mixed models, are typically estimated through maximum likelihood, or variations to this theme, implying that ignorability under MAR can be invoked. This is not the case for non-likelihood marginal models, such as the semi-parametric method of GEEs (Liang and Zeger, 1986), henceforth GEE, which is a second prevalent modelling approach in this area. Such models give valid inferences under the restrictive assumption of MCAR. To be able to analyze the longitudinal non-Gaussian profiles under the weaker MAR assumption, Robins et al. (1995) extended GEEs by using inverse probability weights, resulting in weighted estimating equations, or WGEE. An alternative approach is MI, developed by Rubin (1987). A detailed account is given in Schafer (2003). Missing values are imputed several times, and the resulting complete data sets are analyzed using a standard method, such as GEE. Afterwards, the obtained inferences are combined into a single one (MI-GEE). Regarding the missingness process, standard MI requires MAR to hold, even though extensions exist. Pros and cons of inverse probability weighting methods with respect to MI have been the subject of some debate (the discussion of Scharfstein et al., 1999; Clayton et al., 1998; Carpenter et al., 2006).

In this paper, the focus will be on the comparison between the two GEE versions for incomplete data mentioned above: WGEE and MI-GEE. Comparisons will be made by means of a simulation study, including both small-sample simulations, as well as so-called asymptotic simulations (Rotnitzky and Wypij, 1994). The behavior of both methods in terms of mean squared error (MSE), variance and bias of the estimators will be studied, under correctly specified and misspecified models. In this way, robustness of both methods under misspecification of either the dropout model, the imputation model, or the measurement model, can be explored.

The outline of this paper is as follows. In Section 2, an overview of methods for analyzing incomplete longitudinal non-Gaussian data is given, with main attention on WGEE and MI together with GEE as analysis method. A description of the asymptotic and small-sample simulation design, as well as the results of the simulation study, is provided in Section 3. We conclude with a discussion in Section 4.

2. Methods for incomplete non-Gaussian longitudinal data

Whereas the linear mixed model is seen as a unifying parametric framework for Gaussian repeated measures (Verbeke and Molenberghs, 2000), there are a variety of methods in common use in the non-Gaussian setting.

In line with Fahrmeir and Tutz (2001), Diggle et al. (2002), and Molenberghs and Verbeke (2005), we distinguish between three model families. In a *marginal model*, marginal distributions are used to describe the outcome vector, given a set of predictor variables. The correlation among the components of the outcome vector can then be captured either by adopting a fully parametric model specification or by means of working assumptions, such as in GEE (Liang and Zeger, 1986).