Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 5537–5546
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Covariance miss-specification and the local influence approach in sensitivity analyses of longitudinal data with drop-outs

R.J. O'Hara Hines[a], W.G.S. Hines[b,*]

[a]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ont., Canada N2L 3G1
[b]Department of Mathematics and Statistics, University of Guelph, Guelph, Ont., Canada N1G 2W1

Received 25 May 2006; received in revised form 15 March 2007; accepted 16 March 2007
Available online 22 April 2007

Abstract

Our work examines the performance of proposed local influence diagnostics applied to multivariate normal longitudinal data with drop-outs: these diagnostics prove to be ambiguous as they are sensitive not only to the presence of anomalous records, as intended, but also, unfortunately, to the misspecification of the longitudinal covariance structure of the response. We suggest an unambiguous index for detecting covariance misspecification, and recommend that an analyst use this index first to confirm that the covariance structure is well specified before attempting to interpret the influence diagnostics.

Keywords: Missing responses; Missing at random; Missing-not-at-random; Non-ignorable drop-outs; Local perturbations; Ambiguous influence diagnostics

1. Introduction

There has been considerable interest in modeling longitudinal data with non-ignorable (NI) or, equivalently, missing-not-at-random (MNAR) drop-outs (see, for example, Diggle and Kenward, 1994; Molenberghs et al., 1997; Fitzmaurice and Laird, 2000; Wilkins and Fitzmaurice, 2006). NI drop-outs are records (or, specifically in the biostatistical area, subjects) which disappear from the study prematurely and whose disappearance is related to the subsequently missing measurements. Since the assumptions necessary for fitting, say, a Diggle and Kenward selection model (Diggle and Kenward, 1994) to this type of data are often unverifiable, more recent work has focused on the use of sensitivity analyses (Molenberghs et al., 2001; Verbeke et al., 2001; Jansen et al., 2006). These authors recommend the use of missing-at-random (MAR) modeling supplemented with local influence diagnostics to detect the impact of possible deviations from the MAR assumption. MAR describes drop-outs whose missingness depends only on the observed and not the missing measurements. (Further explanations of the MAR and NI/MNAR drop-out models can be found in Little and Rubin, 1987, Chapter 6.) Jansen et al. (2006) have indicated that these local influence diagnostics should be used 'not to detect individuals that drop-out non-randomly, but rather to detect anomalous subjects that lead to a seemingly MNAR mechanism.' These authors state that 'a careful study of such subjects, combined with appropriate

* Corresponding author.
E-mail address: ghines@uoguelph.ca (W.G.S. Hines).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2007.03.027

treatment (e.g., correction of errors, removal, etc.), can lead to a final MAR model, in which more confidence can be put by the researchers, which ultimately is the goal of every sensitivity analysis.'

There has also been interest in the consequences of incorrectly specifying the covariance structure associated with the longitudinal response (Crowder, 2001; Koreshia and Fang, 2001; Wang and Carey, 2003; Wang and Lin, 2005). These papers consider ways for improving the estimation of the covariance structure, given the loss in regression estimation efficiency when that structure has been misspecified. Our results indicate that one consequence of incorrectly specifying a covariance structure is that the above local influence diagnostics can then be misleading, apparently signaling that particular records are influential when that is not actually the case. To assist analysts, we propose an index for covariance misspecification that we have found to be unambiguous in the presence of influential drop-outs. We recommend that analysts first examine this index to ensure that the covariance structure is well specified before attempting to interpret the influence diagnostics.

To obtain this index, we considered a number of established candidates in a study that was exploratory rather than definitive. The study found that the distributions of all but one of the indices were strongly affected by whether the missingness was MAR or NI, making those indices difficult to use since the determination of the true nature of missingness requires strong and unverifiable assumptions. Our study was pragmatic rather than theoretical, since we are interested in determining methods that perform well in practice as opposed to working well with pre-chosen forms of misspecification.

Clearly, interesting follow-up questions remain about measuring covariance misspecification. Among these is finding a suitable measure of sensitivity to the parameters of the covariance structure. Another is the level of that sensitivity. The two papers by Banerjee and Magnus (1999, 2000) provide sensitivity statistics which measure the effects on the regression coefficient estimates and t- and F-tests when the errors in OLS are, say, from an ARMA(p, q) time series model but are incorrectly assumed to be independent. Since our own interest is with misspecification in any of a potentially large number of ways from a range of possibilities, the challenge of performing comprehensive sensitivity analyses is substantial. Instead, we took the pragmatic approach of selecting a variety of possible test cases that deviated moderately from the structure assumed in the analysis and examining the consequences. Doing so focuses on the existence of weaknesses in the previously proposed influence diagnostics under conditions that might occur in practice, rather than on a deeper analysis of the nature and causes of those weaknesses.

In Section 2, we describe standard models for longitudinal multivariate normal data and for drop-outs that will be used in simulations. The local influence diagnostics and their simulation findings are discussed in Section 3, while the index for detecting covariance misspecification and simulation results are in Section 4. Section 5 provides concluding remarks.

2. Models

One model for longitudinal multivariate normal data uses the regression

$$Y_i = \mu_i = X_i\beta + e_i, \quad i = 1, \ldots, N \tag{1}$$

with $e_i \sim \mathrm{MVN}(0, \mathrm{cov}(Y_i))$, where Y_i is $n_i \times 1$, X_i is $n_i \times p$, β is $p \times 1$ and $\mathrm{cov}(Y_i)$ is $n_i \times n_i$.

A model for drop-out is

$$\mathrm{logit}(\Pr(D_i = j/D_i \geqslant j, y_i)) = \psi_0 + \psi_1 y_{i,j-1} + \omega y_{ij}, \tag{2}$$

where D_i indicates the drop-out time of the ith record y_i. If the coefficient ω is zero then the drop-out model is MAR, otherwise it is NI.

2.1. Simulations details

The models just described can be used for simulation purposes, with suitable choices of parameter values. For the simulation studies reported here, the covariates in X_i in (1) allow for an intercept, 2 groups/treatments coded as -0.5 or 0.5, a linear time trend usually over four observation times coded as -1.5, -0.5, 0.5 or 1.5 and a time by group interaction. The different sets of regression coefficients β used in the simulations are given in Table 1. N varied between 50 and 200 per group. The covariance structure of the generated data over time is either AR(1) or non-AR(1), the latter