Computational Statistics and Data Analysis 56 (2012) 1232–1242

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Generalized estimating equations and regression diagnostics for longitudinal controlled clinical trials: A case study

Maren Vens, Andreas Ziegler *

Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Maria-Goeppert-Strasse 1, 23562 Lübeck, Germany

ARTICLE INFO

Article history:
Available online 23 April 2011

Keywords:
Chronic venous insufficiency
Cook statistic
Deletion diagnostics
Independence estimating equations
Half-normal plot

ABSTRACT

Generalized estimating equations (GEE) were proposed for the analysis of correlated data. They are popular because regression parameters can be consistently estimated even if only the mean structure is correctly specified. GEE have been extended in several ways, including regression diagnostics for outlier detection. However, GEE have rarely been used for analyzing controlled clinical trials. The SB-LOT trial, a double-blind placebo-controlled randomized multicenter trial in which the oedema-protective effect of a vasoactive drug was investigated in patients suffering from chronic insufficiency was re-analyzed using the GEE approach. It is demonstrated that the autoregressive working correlation structure is the most plausible working correlation structure in this study. The effect of the vasoactive drug is a difference in lower leg volume of 2.64 ml per week (p=0.0288, 95% confidence interval 0.27–4.99 ml per week), making a difference of 30 ml at the end of the study. Deletion diagnostics are used for identification of outliers and influential probands. After exclusion of the most influential patients from the analysis, the overall conclusion of the study is not altered. At the same time, the goodness of fit as assessed by half-normal plots increases substantially. In summary, the use of GEE in a longitudinal clinical trial is an alternative to the standard analysis which usually involves only the last follow-up. Both the GEE and the regression diagnostic techniques should accompany the GEE analysis to serve as sensitivity analysis.

1. Introduction

Twenty-five years ago the generalized estimating equations (GEE) for analyzing correlated non-normal data were introduced by Liang and Zeger in a series of papers (see, e.g., Liang and Zeger, 1986; Zeger and Liang, 1986). The strength of this semiparametric approach is that regression coefficients can be consistently estimated in regression models with clustered non-normally dependent variables even if the distribution is partly misspecified. Specifically, only the correct specification of the mean structure is required for consistent estimation. Variances and within-cluster correlations may be misspecified. However, the efficiency of the estimation approach generally depends on the degree of misspecification of the covariance matrix.

The GEE have been extended in several ways, and the extensions include approaches for dealing with missing data (for an overview, see, e.g., Ziegler et al., 2003), approaches for sample size calculations (reviewed in Dahmen and Ziegler, 2004), or regression diagnostics (Preisser and Qaqish, 1996; Ziegler et al., 1995). However, these extensions have rarely been used in applications, partly because of the lack of appropriate software.

* Corresponding author. Tel.: +49 451 500 2781; fax: +49 451 500 2999.
E-mail addresses: maren.vens@imbs.uni-luebeck.de (M. Vens), ziegler@imbs.uni-luebeck.de (A. Ziegler).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.04.010

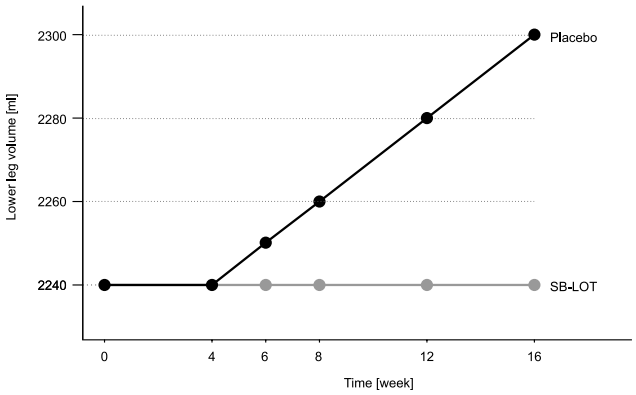

Fig. 1. Expected course of the trial. At randomization (week 0) both patients groups are expected to have identical lower leg volumes. Medical compression stockings are discontinued at week 4, and after week 4 the lower leg volume should increase linearly in the placebo group, while it should remain constant in the SB-LOT group.

The aim of this paper is therefore two-fold. First, we want to illustrate that the application of GEE to a repeated measurement intervention study can be an interesting alternative or at least a supplementation to the standard analysis which only involves the last follow-up and, possibly, adjustments for baseline measurements. Second, we aim at demonstrating that regression diagnostics should supplement the GEE analysis to serve as sensitivity analysis. For illustration, we re-analyze data from a double-blind placebo-controlled randomized multicenter trial, in which the oedema-protective effect of a vasoactive drug was investigated in patients suffering from chronic venous insufficiency after decongestion of the legs. The primary analysis was a baseline-adjusted covariance analysis (ANCOVA) between the two treatment groups (Vanscheidt et al., 2002). A secondary analysis using GEE which aimed at detecting a difference in the slopes will be presented in this paper.

The paper is organized as follows. First, we describe the SB-LOT data (Vanscheidt et al., 2002) which are re-analyzed below. Second, we give a short introduction to GEE, and we briefly discuss approaches for selecting the most plausible correlation structure. Next, we review regression diagnostic methods for GEE, which are primarily based on deletion diagnostics. Results from the re-analysis of the SB-LOT data are presented, and findings from regression diagnostics are displayed. We specifically show for this data set that the removal of outliers does not alter the overall conclusion of the study. However, the goodness of fit as assessed by half-normal plots and simulated envelopes improves.

2. The SB-LOT data

For illustration we use a parallel group design with repeated measurements. In this double-blind placebo-controlled randomized multicenter trial, the oedema-protective effect of a vasoactive drug was investigated in patients suffering from chronic venous insufficiency after decongestion of the legs (Vanscheidt et al., 2002). At the baseline, 226 patients were randomized to medical compression stockings plus SB-LOT (90 mg Coumarin and 540 mg Troxerutin per day) or medical compression stockings plus placebo for the first 4 weeks and SB-LOT or placebo for the following 12 weeks of the study. In the first four weeks all patients wore medical compression stockings. At the first follow-up, i.e., four weeks after randomization, the stockings were discontinued in both treatment groups. Subsequently, the investigators expected an increase in the lower leg volume at subsequent follow-ups in the placebo group, while the lower leg volume was expected to be constant in the active treatment group SB-LOT (Fig. 1).

Patients were followed-up five times: 4, 6, 8, 12, and 16 weeks after initiation of the drug therapy. Thus, follow-ups were not equally spaced. The primary efficacy endpoint was the lower leg volume measured by water plethysmometry. The primary analysis was a baseline-adjusted (visit 0) covariance analysis (ANCOVA) of the difference of the leg volume at the final visit minus the volume at baseline, i.e., the time point of randomization for demonstrating a difference of the vasoactive drug when compared with placebo. The intention to treat (ITT) group of this trial consists in 113 patients per treatment group. Table 1 displays the lower leg volume (ml) in the course of the trial for these patients. As expected, the correlation between observations at different time points is high (Table 2).

The secondary analysis which will be presented here aims at detecting a difference in the slopes by making use of the repeatedness nature of the data. Because the correlation between the measurements at different time points is considered nuisance, this re-analysis represents a typical setting for a GEE1 analysis.

3. Generalized estimating equations

Let n be the number of independent clusters $i = 1, \ldots, n$, and, for simplicity, assume that there are T observations per cluster ($t = 1, \ldots, T$). For each dependent variable y_{it} a p-dimensional vector of independent variables $\mathbf{x}_{it}$ is available. Data are collected in column vectors $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$ and $T \times p$ dimensional matrices $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})'$.