Ann. of Appl. Math.
32:2(2016), 200-220

EMPIRICAL LIKELIHOOD APPROACH FOR LONGITUDINAL DATA WITH MISSING VALUES AND TIME-DEPENDENT COVARIATES[*][†]

Yan Zhang, Weiping Zhang[‡], Xiao Guo

(*Dept. of Statistics and Finance, University of Science and Technology of China, Anhui 230026, PR China*)

Abstract

Missing data and time-dependent covariates often arise simultaneously in longitudinal studies, and directly applying classical approaches may result in a loss of efficiency and biased estimates. To deal with this problem, we propose weighted corrected estimating equations under the missing at random mechanism, followed by developing a shrinkage empirical likelihood estimation approach for the parameters of interest when time-dependent covariates are present. Such procedure improves efficiency over generalized estimation equations approach with working independent assumption, via combining the independent estimating equations and the extracted additional information from the estimating equations that are excluded by the independence assumption. The contribution from the remaining estimating equations is weighted according to the likelihood of each equation being a consistent estimating equation and the information it carries. We show that the estimators are asymptotically normally distributed and the empirical likelihood ratio statistic and its profile counterpart follow central chi-square distributions asymptotically when evaluated at the true parameter. The practical performance of our approach is demonstrated through numerical simulations and data analysis.

Keywords empirical likelihood; estimating equations; longitudinal data; missing at random

2000 Mathematics Subject Classification 62G05

1 Introduction

Longitudinal data frequently occur in many studies such as medical follow-up studies. A key characteristic of longitudinal data is that outcomes measured repeatedly on the same subject are typically correlated. Regression methods for such

[*] This work was supported by the NNSF of China (No.11271347) and the Fundamental Research Funds for the Central Universities.

[†] Manuscript received September 25, 2015; Revised March 31, 2016

[‡] Corresponding author. E-mail: zwp@ustc.edu.cn

datasets accounting for within-subject correlation is abundant in the literatures [3,4]. Among which, the generalized estimating equations (GEEs) method by [12] has been widely used since it considers the mean structure and the correlation structure separately. When the marginal mean outcome given the covariates at current time is the same as that on all the past, present and future covariate values, the GEEs method assures consistency of the mean estimates even if the correlation is misspecified, and achieves efficiency if the correlation is correctly specified.

In practice, however, it is common that some covariates may vary over time in a longitudinal study, that is, some of the covariates may be time-dependent. For example, in the Mother's Stress and Children's Morbidity Study (MSCM, [1]), the daily ratings of child illness Y_{it} and maternal stress X_{it} are measured during a 28-day follow-up period. Obviously, Y_{it} and X_{it} are time-dependent, both of which vary over time and may correlate with the other measurements. It has been noted that the consistency of GEEs is not assured with arbitrary working correlation structures when there are time-dependent covariates [17]. The reason is that the estimating functions generated by the longitudinal data are no longer unbiased under an arbitrary correlation structure unless the marginal mean outcome given the covariates at the current time is the same as that on all the past, present and future covariate values. Clearly, the use of an independence assumption guarantees the consistency of GEEs but can result in a substantial loss in efficiency due to the fact that, only a subset of all the unbiased estimating functions is used. Some methods have been developed for such situations, for example, [10] classified the time-dependent covariates into three different types based on the moment conditions that are valid to the covariates. They then introduced a generalized method of moments (GMM) to combine all available valid estimating equations optimally. In general, the basic idea is to select estimating functions by minimizing some criteria, see [9,15,24] among others. Recently, [10] proposed a shrinkage empirical likelihood (EL, [16]) approach by including all the estimating functions under the independence correlation assumption and/or those are known to be unbiased a priori, and shrinking all other estimating functions according to the likelihood of each being a biased, uninformative or informative estimating equation. Their approach avoids identifying the uninformative and biased estimating functions and allows different shrinkage parameters for different estimating functions.

Another common problem in the longitudinal data analysis is the missing data problem. For example, there were approximately 4% of dropout in the illness record during the 4 weeks in MSCM study. Statistical methods are available to handle such issue by incorporating missing data mechanisms to provide valid statistical inference, including complete-case analysis method, imputation method, inverse