Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 50 (2006) 2702–2714
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Analysis of longitudinal data with intermittent missing values using the stochastic EM algorithm

Ahmed M. Gad[a,*], Abeer S. Ahmed[b]

[a]Department of Statistics, Faculty of Economics and Political Science, Cairo University, Cairo, Egypt
[b]The National Centre for Social and Criminological Research, Cairo, Egypt

Received 26 May 2004; received in revised form 10 April 2005; accepted 10 April 2005
Available online 3 May 2005

Abstract

Longitudinal data are not uncommon in many disciplines where repeated measurements on a response variable are collected for all subjects. Some intended measurements may not be available for some subjects resulting in a missing data pattern. Dropout pattern occurs when some subjects leave the study prematurely. The missing data pattern is defined as intermittent if a missing value followed by an observed value. When the probability of missingness depends on the missing value, and may be on the observed values, the missing data mechanism is termed as nonrandom. Ignoring the missing values in this case leads to biased inferences. The stochastic EM (SEM) algorithm is proposed and developed to find parameters estimates in the presence of intermittent missing values. Also, in this setting, the Monte Carlo method is developed to find the standard errors of parameters estimates. Finally, the proposed techniques are applied to a real data from the International Breast Cancer Study Group.

Keywords: Repeated measures; Nonrandom intermittent missing; The stochastic EM algorithm; Standard errors; Quality of life; Breast cancer

1. Introduction

Longitudinal studies play a prominent role in many disciplines such as medicine, public health and social sciences. In longitudinal studies, each subject is measured repeatedly

* Corresponding author. Quantitative Methods Department, College of Administrative Sciences, King Saud University, Riyadh 11451, P.O. Box 2459, Kingdom of Saudi Arabia. Tel.: +966 1 4674081; fax: +966 1 4674091.
E-mail addresses: hagas10@hotmail.com, shyba2@yahoo.co.uk (A.M. Gad).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2005.04.006

for the same response (outcome) variable either under different conditions, or at different times, or both. The response variable may be continuous, categorical or ordinal. The main interest of longitudinal studies, usually, is to study the change in the response variable over time. Longitudinal data analysis requires special statistical methods due to inter-correlation between the observations of the same subject.

The missing values is a common phenomenon in longitudinal studies. The missing data occur whenever, one or more of, measurement sequences are incomplete. A distinguishing feature of incomplete longitudinal data analysis is the need to address the underlying causes of missing values. If the missing data process depends on the missing values themselves, the missing data mechanism is nonrandom. Ignoring the missing values with such data would lead to biased conclusion. Another important feature is whether the missing values pattern is dropout (monotone) or intermittent (nonmonotone). The dropout pattern, in the sense that some subjects may withdraw prematurely, i.e. any missing value is never followed by an observed value. The intermittent pattern whenever an observed value is available even after a missing value occurs. Missing data mechanism is classified to three different types due to Rubin (1976) and Little and Rubin (1987). These types are missing completely at random, missing at random, and nonrandom (informative) missingness. Nonrandom dropouts in longitudinal studies have been extensively studied, but there has been less attention paid to nonrandom intermittent missing values.

The maximum-likelihood estimates for incomplete data can be obtained using the EM algorithm (Dempster et al., 1977). However, in the nonrandom case, the simplicity of the EM algorithm is lost. The expectation step is problematic and does not admit a closed form solution. Also, in some situations, the M-step is computationally unattractive. Many authors have tried to introduce new variants of the EM algorithm that can overcome the complexity of either the E-step or the M-step. Several approaches are available to facilitate the M-step when it is not simple, see for example, Lange (1995), Liu and Rubin (1994), Meng and Rubin (1993), Meng and van Dyk (1995) and Rai and Matthews (1993). A possible solution for the intractable E-step is to use the Monte Carlo EM algorithm (Tanner and Wong, 1987; Wei and Tanner, 1990) and a stochastic version of the EM algorithm (Celuex and Diebolt, 1985; Delyon et al., 1999; Diebolt and Ip, 1996; Gu and Kong, 1998; Zhu and Lee, 2002). A relatively recent review of the EM algorithm and its extensions is in McLachlan and Krishnan (1997) and references therein. The stochastic EM (SEM) algorithm is a stochastic version of the EM algorithm, which has been introduced by Celuex and Diebolt (1985), and subsequently in Diebolt and Ip (1996), as a way for executing the E-step using simulation.

The EM algorithm does not provide directly the standard errors of the estimates. Hence, methods for evaluating these standard errors need to be considered. Several methods have been introduced to solve this problem, see for example, Louis (1982), Meilijson (1989) and Meng and Rubin (1991). Efron (1994) and Ip (1994) have introduced a stochastic version of the Louis' method (the Monte Carlo method).

Diggle and Kenward (1994) propose a modelling framework for longitudinal data with nonrandom dropout. This approach relates the probability of dropout at any time with the current observation and the previous observations. The estimation process is cumbersome due to integrations that need to obtained numerically.

Troxel et al. (1998) generalize the Diggle–Kenward model to handle intermittent missing data pattern. They use the Nelder–Mead simplex algorithm (Nelder and Mead, 1965) to