Journal of Multivariate Analysis 166 (2018) 62–77

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Analysis of ordinal longitudinal data under nonignorable missingness and misreporting: An application to Alzheimer's disease study

Subrata Rana [a], Surupa Roy [b], Kalyan Das [c,*], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Department of Statistics, University of Calcutta, Kolkata 700019, India
[b] Department of Statistics, St. Xavier's College, Kolkata 700016, India
[c] Department of Mathematics, IIT Bombay, Powai, Mumbai 400076, India

ARTICLE INFO

Article history:
Received 22 June 2017
Available online 26 February 2018

AMS subject classifications:
62-07
62F99

Keywords:
Bivariate binary model
MCNREM
Miscategorization
Missing data
Selection model

ABSTRACT

In many epidemiological and clinical studies, observations on individuals are recorded longitudinally on a Likert-type scale. In the process of recording, or due to some other causes, a proportion of outcomes and time-dependent covariates may be missing in one or more follow-up visits (non monotone missing). Even when the number of patients with intermittent missing data is small, exclusion of those patients from the study seems unsatisfactory. This apart, often due to misreporting, miscategorization of response can occur that results in potentially invalid inference when no correction is made. We propose a joint mixed model that corrects the likelihood function to account for missing response and/or covariates and adjusts the likelihood to tackle miscategorization of response. Under this extreme complex but useful setup, we seek to estimate the parameters of the proposed model that accounts for baseline and/or time dependent covariates. Monte Carlo expectation–maximization (MCEM) is a convenient approach for estimating the parameters in the model. A simulation study was carried out to assess the approach. We also analyzed Alzheimer's Disease Neuroimaging Initiative (ADNI) data where some responses and covariates are missing and some responses are possibly miscategorized. Our investigation reveals that apolipo-protein plays a significant role in Alzheimer's disease progression. This was not visible in earlier analyses of ADNI data.

1. Introduction

In many research problems related to biological, social and medical sciences, multivariate data arise from repeated measurements on a sample of subjects over time. To analyze such longitudinal data, one must consider the relationship between the serial observations made on a given unit and hence it is inappropriate to use a general multivariate model

* Corresponding author.
E-mail address: kalyan@math.iitb.ac.in (K. Das).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.jmva.2018.02.004
0047-259X/

for studying covariate effects. A two-stage random-effect model [12] or a generalized linear model [17] can be used in this situation. Often, the response of interest is measured on an ordinal scale with more than two categories.

Methods for longitudinal ordinal data analysis have been actively pursued in the past; see, e.g., [13,15,20]. A marginalized latent variable model could be used to analyze data for this situation. A marginalized model separates the systematic variation in the data from random variation. Basically it uses two models in conjugation. One is the mean or regression model for the ordinal responses and the other is the dependence model. A cumulative logit model with proportional odds assumption is commonly used as the mean model. The covariates used in the regression model may include both qualitative and quantitative variables, some of which are measured at the baseline visit while others vary over time. In the dependence model, additional variables are introduced as random components to structure the relation between repeated measurements. Such a modeling approach was used by Heagerty [8] in the context of binary responses. Lee and Daniels [13] extended the marginalized latent variable model to accommodate ordinal responses.

The modeling of longitudinal responses through latent variables becomes complex when data on the response and some covariates go missing. A simple solution is to ignore the missing observations and perform a complete-case analysis. However this leads to inefficient inference, especially when the missing mechanism is non-ignorable. Following Little and Rubin [19], there are three types of missing data processes. Data are said to be missing completely at random (MCAR) if the missing data process does not depend on missing or observed data, and the process is said to be missing at random (MAR) if the missing data process depends on the observed data only.

In this paper we consider a non-ignorable missing data mechanism (MNAR) in which the missing data process depends on both observed and unobserved data. The MNAR pattern of missing data has been considered by Ibrahim and Lipsitz [10] in the context of generalized linear models when covariates are missing while Troxel et al. [30] and Ibrahim et al. [9] have considered missing responses. Most of the work on missing data focuses on either missing response or covariate. Stubbendick and Ibrahim [28,29] adopted a maximum likelihood approach for estimating the model parameters in a longitudinal study when both response and covariates are missing, but in their case the response belonged to the exponential family. Chen et al. [2] have considered missing response and covariate in the context of longitudinal binary data when the missing data mechanism is MAR. Here we have considered non-ignorable missingness in response as well as in a covariate.

Missing data analysis in a longitudinal set up is usually carried out under the assumption that the levels of the ordinal response are correctly classified. In medical or social sciences, however, the true level of the response is often not identified correctly. The reason for this misclassification may be misreporting by a subject or faulty diagnostic tests. For example, in a disease progression longitudinal study, there may be misreporting about a patient's disease severity at subsequent visits if the ordinal responses are collected and maintained by semi-experts. Also in the job characteristic data [25], employees were asked to respond to different aspects of the job which was measured on a five-point ordinal scale ranging from strongly agree to strongly disagree. The employees' response is supposed to be misclassified.

Miscategorization of categorical data has been considered by many researchers [4,26]. Espeland and Hui [7], Buonaccorsi [1], Pepe [24] considered a double sampling procedure to obtain the estimates of misclassification rates for discrete data, binary data and continuous data, respectively. However, only few studies are available in the literature on ordinal categorical data. Eickhoff and Amemiya [6] considered a known monotone misclassification pattern in either direction for polytomous outcome variable. Poon and Wang [25] discussed the use of a surrogate variable while modeling multivariate ordinal responses. Chen et al. [3] considered a generalized estimating equation approach while dealing with error prone ordinal responses and covariates.

In this paper we develop a flexible model to account for missing response and covariate which is also adjusted for misclassification of the observed ordinal response. The study, which is carried out under a longitudinal set up, incorporates the dynamic nature of the missingness pattern. However, it is assumed that the misclassification pattern remains the same over the visits.

The paper is organized as follows. In Section 2, we propose a flexible model and discuss the identifiability issues related with the model parameters. Section 3 describes the estimation methodology implemented via a Monte Carlo Newton–Raphson Expectation Maximization method. A simulation study is reported in Section 4 to assess the approach. In Section 5, Alzheimer's Disease Neuroimaging Initiative (ADNI) data are analyzed and finally, we conclude with some observations in Section 6.

2. Model formulation

2.1. The response process

Consider a study involving N subjects in which subject $i \in \{1, \ldots, N\}$ is assessed on $n_i \leq T$ occasions. Let $Y_i = (y_{i1}, \ldots, y_{in_i})^\top$ be a vector of L-category ordinal responses for the ith subject. Let also $Z_i = (z_{i1}, \ldots, z_{iq})^\top$ be a vector of baseline covariates and $\widetilde{X}_{i1}, \ldots, \widetilde{X}_{ip}$ be p time-varying covariate vectors, where $\widetilde{X}_{ij} = (x_{i1j}, \ldots, x_{itj}, \ldots, x_{in_ij})^\top$ for each $j \in \{1, \ldots, p\}$. We denote the $n_i \times p$ matrix of p-time varying covariates for the ith subject by $X_i = (\widetilde{X}_{i1}, \ldots, \widetilde{X}_{ip})$. The tth row of the matrix X_i contains the subject-specific time-dependent covariates which we denote by $X_{it} = (x_{it1}, \ldots, x_{itp})^\top$.

To start with, we assume that the responses and the covariates corresponding to each subject are completely observable on all occasions. We consider a marginalized latent variable model for the regression set up of the longitudinal ordinal response. Two separate models are considered in conjugation: the mean (or regression) model and the dependence model.