Statistics and Probability Letters 113 (2016) 16–22

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

A weighted simulation-based estimator for incomplete longitudinal data models

Daniel H. Li [a], Liqun Wang [b,*]

[a] Juno Therapeutics, Seattle, WA, USA

[b] University of Manitoba, Winnipeg, Canada

ARTICLE INFO

Article history:
Received 25 October 2015
Received in revised form 8 February 2016
Accepted 11 February 2016
Available online 23 February 2016

Keywords:
Generalized linear mixed model
Inverse probability weighting
Missing data
Simulation-based estimator

ABSTRACT

Recently, Li and Wang (2012a,b) and Wang (2007) have proposed a simulation-based estimator for generalized linear and nonlinear mixed models with complete longitudinal data. This estimator is constructed using the simulation-by-parts technique which leads to the unique feature that it is consistent even using finite number of simulated random points. This paper extends the methodology to deal with incomplete longitudinal data by applying the inverse probability weighting method for the monotone missing-at-random response data. The finite sample performance of this estimator is investigated through simulation studies and compared with the multiple imputation approach.

1. Introduction

In biomedical, environmental and social sciences research, longitudinal data analysis is widely used and constitutes the fundamental statistical research methodologies. Generalized linear mixed models (GLMM) have been widely used in the modeling of longitudinal data. Li and Wang (2012a) proposed a simulation-based estimator (SBE) for GLMM based on the first two marginal moments of the response variables, which does not rely on the normality distribution assumption for random effects. Li and Wang (2012b) extended the SBE to the GLMM where some covariates are measured with error. This approach was originally studied by Wang (2007) for nonlinear mixed effects models. The SBE is constructed using a novel simulation-by-parts technique to ensure its consistency by using finite number of simulated random points. This is the key difference from many other simulation-based estimators proposed in the literature, where they require the number of simulated random points go to infinity to achieve consistency. So far, the SBE is only studied under complete data settings although incomplete or missing data are common in longitudinal studies. For example, in clinical trials, missing data are almost inevitable because subjects may decide to withdraw from the study at anytime prior to completion or subjects are not compliant to protocol for scheduled assessments. Problems arise if the mechanism leading to the missing data depends on the response process. It is known that ignoring missing data and using naive methods may introduce bias, reduce the power of inference and lead to misleading conclusions (Little and Rubin, 2002).

The extension of the SBE to account for incomplete longitudinal data is non-trivial and needs to be addressed to allow this estimator used in more general settings. In this paper, we discuss the validity of SBE under different missing data mechanism and modify it for the data missing at random (MAR) with monotone missingness through the inverse probability weighting (IPW) method. The IPW is a general methodology for constructing parameter estimators in semi-parametric models with complete as well as missing data (Robins et al., 1995; Yi and Cook, 2002). Another popular approach to deal with missing

* Correspondence to: Department of Statistics, University of Manitoba, 186 Dysart Road, Winnipeg, Manitoba, Canada R3T 2N2.
E-mail address: liqun.wang@umanitoba.ca (L. Wang).

http://dx.doi.org/10.1016/j.spl.2016.02.004
0167-7152/

data is the multiple imputation (Rubin, 1987; Schafer, 1997). We also investigate the performance of the SBE using this strategy.

The structure of the paper is as follows. In Section 2, we introduce and review the missing data mechanism, pattern and estimation. Section 3 provides details on the proposed weighted simulation-based estimator, and addresses some practical computational issues. In particular, Section 3.2 handles the data missing completely at random (MCAR), while Section 3.3 focuses on the data MAR. Some simulation studies are conducted in Section 4 to examine the finite sample performance of the proposed estimator, and concluding remarks are given in Section 5.

2. Missing data framework and notation

2.1. Missing data mechanism

To obtain valid inferences, it is essential to consider the reason for missingness. Let Y_{ij} be the jth response for the ith subject, $R_i = (r_{i1}, r_{i2}, \ldots, r_{in})'$ be the vector of missing data indicators for $Y_i = (y_{i1}, \ldots, y_{in})'$, such that $r_{ij} = 1$ if response y_{ij} is observed, and 0 otherwise. We partition Y_i into Y_i^O and Y_i^M, where Y_i^O contains those y_{ij} for which $r_{ij} = 1$ and Y_i^M contains the remaining components. Assuming $X_i = (x_{i1}, \ldots, x_{in})'$ to be a vector of covariates always observed, Little and Rubin (2002) classified missing data mechanism into three types: (1) MCAR, where the missingness is unrelated to the responses so that $P(R_i|Y_i, X_i) = P(R_i|X_i)$. (2) MAR, where the missingness depends only on the observed responses so that $P(R_i|Y_i, X_i) = P(R_i|Y_i^O, X_i)$. This is a weaker and more plausible assumption than MCAR. (3) MNAR, where the missingness depends on both observed and unobserved responses.

2.2. Missing data patterns

There are two broad classes of missing data patterns: intermittent missing and dropout. Intermittent missing pattern refers to the scenario that a subject completes the study but skips a few occasions in the middle of the study period. Dropout (attrition, lost of follow-up) is a particular example of monotone pattern of missingness, which means if one observation is missing, then all subsequent observations are unobserved. Intermittent missing is often easier to deal with because the subject is still participating in the study and the reason of missing values can be ascertained. Dropout is more serious because the subject is no longer available and it is not certain whether the dropout is related to the observed or unobserved outcome. MAR mechanisms are commonly assumed when the interest lies on the parameter estimation (Robins et al., 1995; Lindsey, 2000).

2.3. Estimation of missing data process

Let $\lambda_{ij} = P(r_{ij} = 1|r_{i,j-1} = 1, X_i, Y_i^O)$ be the conditional probability that subject i is observed at time j, given that the subject is present at time $j-1$; and $\pi_{ij} = P(r_{ij} = 1|X_i, Y_i^O)$ be the marginal probability that subject i is present at time j. Then $\pi_{ij} = \prod_{t=2}^{j} \lambda_{it}$. Generally it is assumed that all individuals are observed on the first occasion so that $r_{i1} = \lambda_{i1} = 1$. Further, let $\pi_{ijk} = P(r_{ij} = 1, r_{ik} = 1|X_i, Y_i^O)$ be the probability of observing both y_{ij} and y_{ik} given the response history and covariates. Usually λ_{ij} is estimated using a logistic regression model $\text{logit}\lambda_{ij} = A_{ij}'\alpha$, where A_{ij} is a vector consisting of information on X_i and response history, and α is the vector of parameters (Diggle and Kenward, 1994; Fitzmaurice et al., 1996; Molenberghs et al., 1997; Yi and Cook, 2002).

3. Weighted simulation-based estimator

3.1. GLMM formulation

Suppose subject i is measured repeatedly on n_i occasions. In a GLMM it is assumed that, given the covariates and random effects $b_i \in \mathbb{R}^q$, the responses y_{ij} are conditionally independent and have distribution from the exponential family

$$f(y_{ij}|b_i, X_i, Z_i) = \exp\left\{\frac{\omega_{ij}y_{ij} - a(\omega_{ij})}{\phi} + c(y_{ij}, \phi)\right\}, \quad i = 1, \ldots, N,\ j = 1, \ldots, n_i, \tag{3.1}$$

where ϕ is a dispersion parameter, ω_{ij} is the canonical parameter and $a(\cdot)$ and $c(\cdot)$ are known functions. The conditional mean and variance

$$\mu_{ij}^c = E(y_{ij}|b_i, X_i, Z_i) = a^{(1)}(\omega_{ij}) \tag{3.2}$$

$$v_{ij}^c = \text{Var}(y_{ij}|b_i, X_i, Z_i) = \phi a^{(2)}(\omega_{ij}) \tag{3.3}$$