Journal of Multivariate Analysis 135 (2015) 59–70

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Achieving semiparametric efficiency bound in longitudinal data analysis with dropouts

Peisong Han[a,*], Peter X.-K. Song[b], Lu Wang[b]

[a] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[b] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

ARTICLE INFO

Article history:
Received 2 February 2014
Available online 15 December 2014

AMS subject classifications:
62F10
62F12

Keywords:
Augmented inverse probability weighting (AIPW)
Double robustness
Empirical likelihood
Local efficiency
Missing at random

ABSTRACT

In longitudinal data analysis with dropouts, despite its local efficiency in theory, the augmented inverse probability weighted (AIPW) estimator hardly achieves the semiparametric efficiency bound in practice, even if the variance–covariance of the longitudinal outcomes is correctly modeled. In this paper, we propose a method based on conditional empirical likelihood. Assuming missing at random (MAR) mechanism, our estimator is doubly robust and locally efficient. Unlike the AIPW estimator, our estimator does not require to model any second moments, including the variance–covariance of the longitudinal outcomes, in order to achieve the semiparametric efficiency bound.

1. Introduction

In longitudinal studies, repeated measurements are collected from the subjects over certain time period. Dropouts are commonly seen in longitudinal studies, where dropout means that some subject leaves the study in the middle of the follow-up and does not return. The missing data caused by dropouts often complicate statistical estimation and inference. Unless the dropouts are completely at random [11,12], analysis based on a direct application of the generalized estimating equations (GEE) method [10] leads to biased estimation.

To correct for the selection bias due to dropouts, Robins et al. [22] and Robins and Rotnitzky [20] proposed the inverse probability weighted (IPW) GEE method. Under the assumption of missing at random (MAR) [12], their estimator is consistent if the missingness probabilities are correctly modeled. According to Robins et al. [21], an augmentation term that extracts more information from subjects with incomplete measurements can be incorporated to improve estimation efficiency. Along this line, Tsiatis [30] presented a detailed study of the augmented inverse probability weighted (AIPW) complete-case GEE method. In addition to potential efficiency gain, this method yields an estimator that is doubly robust [26], in the sense that the estimator is consistent if either the missingness probabilities or the conditional expectations of certain functions of the full data given the observed data at each level of missingness are correctly modeled. For more discussion on the AIPW method and double robustness, please also refer to Robins et al. [21], Rotnitzky and Robins [25], van der Laan and Robins [32], Bang and Robins [1], Rotnitzky [23], Seaman and Copas [27], Tsiatis et al. [31], and Rotnitzky et al. [24].

[*] Corresponding author.
E-mail addresses: peisonghan@uwaterloo.ca (P. Han), pxsong@umich.edu (P.X.-K. Song), luwang@umich.edu (L. Wang).

http://dx.doi.org/10.1016/j.jmva.2014.12.006

For longitudinal data with dropouts, under the semiparametric model defined by (i) the conditional mean structure of the longitudinal outcomes given the covariates and (ii) the missing at random mechanism, Robins and Rotnitzky [20] derived the efficiency bound, which is the highest level of estimation efficiency achievable by any regular and asymptotically linear estimator under this semiparametric model. Obtaining an estimator that achieves the efficiency bound is not easy. One such success was given by Robins and Rotnitzky [20], who proposed to model the following quantities: (i) the missingness probabilities, (ii) the conditional expectations of the longitudinal outcomes given the observed data at each level of missingness, and (iii) certain second moments of the underlying data distribution (more precisely, the variance–covariance of the vector defined by (4) in Section 3). Their estimator achieves the efficiency bound when all those quantities are correctly modeled, and thus is locally efficient. Refer to Tsiatis [30] and Rotnitzky [23] for more details on how to obtain locally efficient estimators in longitudinal data analysis with dropouts. However, the second moments required by Robins and Rotnitzky [20] are not simply the variance–covariance of the longitudinal outcomes. Therefore, correctly modeling the variance–covariance of the longitudinal outcomes does not make an existing estimator achieve the efficiency bound. As a matter of fact, the second moments required to achieve the efficiency bound may be very difficult to model in practice due to their complex forms and the unknown data distribution. Hence, a method that avoids modeling any second moments yet still achieves the efficiency bound is appealing.

In recent literature, combined with the AIPW approach, empirical likelihood [14,15,17] has become a powerful tool in solving missing data problems. The double robustness property can be preserved [28,29,19,18] and even generalized [6,4,5]. The application of the empirical likelihood method relies on a set of estimating functions. In regression analysis, to achieve high efficiency, the construction of the optimal estimating functions requires to correctly model certain second moments of the data distribution (e.g., [2]). To avoid modeling the second moments, in this paper, we propose a method based on conditional empirical likelihood [35,9], a variant of empirical likelihood for models defined by conditional moment restrictions. The proposed method does not need to model any second moments, including the variance–covariance of the longitudinal outcomes. The resulting estimator is doubly robust; that is, the estimator is consistent if either the missingness probabilities or the conditional expectations of the longitudinal outcomes given the observed data at each level of missingness are correctly modeled. When both quantities are correctly modeled, our estimator achieves the semiparametric efficiency bound.

This paper is organized as follows. Necessary notations and the description of the data and model are presented in Section 2. Section 3 introduces the proposed method. Section 4 focuses on the numerical implementation. Large sample properties of the proposed estimator are studied in Section 5. Section 6 covers the simulation experiments. Some relevant discussions are given in Section 7. The Appendix consists of some technical details.

2. Data and model

Let Y_{ik} and $\boldsymbol{X}_{ik}$ denote the outcome and a vector of covariates collected from subject i ($i = 1, \dots, N$) at time k ($k = 0, \dots, K$), respectively, where time 0 denotes the baseline. In many practical studies, a set of auxiliary variables $\boldsymbol{S}_{ik}$ may also be collected at each visit k. Although they are not of direct statistical interest, these auxiliary variables can usually help explain the missingness mechanism and improve estimation efficiency. Therefore, our development in this paper takes their possible presence into account. Write $\boldsymbol{Y} = (Y_{0,} \dots, Y_K)^{\mathrm{T}}$, $\boldsymbol{X} = (\boldsymbol{X}_0^{\mathrm{T}}, \dots, \boldsymbol{X}_K^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{S} = (\boldsymbol{S}_0^{\mathrm{T}}, \dots, \boldsymbol{S}_K^{\mathrm{T}})^{\mathrm{T}}$. Our interest is to estimate the unknown p-dimensional vector $\boldsymbol{\beta}_0$ in the following mean regression model:

$$E(Y_k \mid \boldsymbol{X}) = \mu_k(\boldsymbol{X}, \boldsymbol{\beta}_0) \quad (k = 1, \dots, K), \tag{1}$$

where μ_k are user-specified link functions depending on the nature of the outcome. For example, the identity link may be used for continuous outcome, and the logit link may be used for binary outcome. The relationship between $\boldsymbol{Y}$ and $\boldsymbol{S}$ is not of direct interest, and thus $\boldsymbol{S}$ is not included in the regression model (1).

To account for possible dropouts, define R_{ik} to be the indicator of observing subject i at time k; that is, $R_{ik} = 1$ if subject i is still in the study at time k, and $R_{ik} = 0$ otherwise. Without loss of generality, assume that data at the baseline are always observed; that is, $R_{i0} = 1$. Due to the fact that dropouts lead to monotone missingness, we have that $R_{ik} = 0$ implies $R_{i(k+1)} = 0$ ($k = 1, \dots, K-1$). Write $\boldsymbol{R} = (R_{0,} \dots, R_K)^{\mathrm{T}}$. In this paper, we allow the auxiliary variables to be missing together with the outcome, but assume the covariates are fully observed. Such scenario occurs, for example, when the covariates are external time-dependent variables or deterministic functions of time and baseline covariates. Therefore, our observed data are N independent and identically distributed copies of $(\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{R}^{\mathrm{T}}, \boldsymbol{R}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}, \boldsymbol{R}^{\mathrm{T}}\boldsymbol{S}^{\mathrm{T}})^{\mathrm{T}}$. The missing data caused by dropouts are assumed to be missing at random, in the sense that for any $k = 1, \dots, K$,

$$P(R_k = 1 \mid R_{k-1} = 1, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S}) = P(R_k = 1 \mid R_{k-1} = 1, \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1}), \tag{2}$$

where $\bar{\boldsymbol{Y}}_{k-1} = (Y_0, \dots, Y_{k-1})^{\mathrm{T}}$ and $\bar{\boldsymbol{S}}_{k-1} = (\boldsymbol{S}_0^{\mathrm{T}}, \dots, \boldsymbol{S}_{k-1}^{\mathrm{T}})^{\mathrm{T}}$. In other words, the probability of observing a subject at the current scheduled visit, given the fact that the subject was observed at the previous visit, does not depend on the current or future unobserved data, but only depends on the observed history. Denote the probability in (2) by $\pi_k(\boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1})$. As usual, the probability of observing the complete data is assumed to be bounded away from zero, or equivalently,

$$\pi_k = \pi_k(\boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1}) > c > 0 \quad (k = 1, \dots, K) \tag{3}$$

for some constant c. The semiparametric model for longitudinal data with dropouts considered in this paper is defined by (1)–(3). This model is among the semiparametric models considered by Robins et al. [22] and Robins and Rotnitzky [20].