Computational Statistics and Data Analysis 121 (2018) 104–112

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Analysis of longitudinal data with covariate measurement error and missing responses: An improved unbiased estimating equation

Huiming Lin [a,b], Guoyou Qin [a,b,*], Jiajia Zhang [c], Zhongyi Zhu [d]

[a] Department of Biostatistics, School of Public Health and Key Laboratory of Public Health Safety, Fudan University, Shanghai, 200032, China
[b] Collaborative Innovation Center of Social Risks Governance in Health, China
[c] Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC 29208, USA
[d] Department of Statistics, Fudan University, Shanghai, 200433, China

ARTICLE INFO

Article history:
Received 6 July 2017
Received in revised form 25 November 2017
Accepted 25 November 2017
Available online 26 December 2017

Keywords:
Marginal method
Measurement error
Missing data
Partially linear models
Robustness

ABSTRACT

Because of the data collection process, measurement error and missing responses are common in longitudinal data, and correctly addressing these scenarios becomes one of main challenges in longitudinal data analysis. First, an unbiased estimating equation is proposed to improve the efficiency of parameter estimations for the marginal mean model for longitudinal data with covariate measurement error. The proposed unbiased estimating equation is asymptotically more efficient than the method in Qin et al. (2016a). Second, the proposed method can be extended to handle more complicated scenarios. Specifically, robust estimation for partially linear models with missing responses and covariate measurement error is considered. The proposed robust estimation does not require specifying the distribution of the covariate or the measurement error and is computationally easy to implement. Simulation studies are conducted to evaluate the improvement of the proposed method over existing methods (Qin et al., 2016b), and a sketch of the proof of its asymptotic property is provided. Finally, the proposed method is applied to the data from the Lifestyle Education for Activity and Nutrition (LEAN) study.

1. Introduction

Longitudinal studies are commonly conducted in biomedical and epidemiological research, and mis-measured covariates are typical in longitudinal data for various reasons, such as the equipments' accuracy or the testing personnel's skills. For example, the Lifestyle Education for Activity and Nutrition (LEAN) study (Barry et al., 2011), conducted at the University of South Carolina, Columbia, SC, USA, was designed to determine the effectiveness of an intervention to enhance weight loss over a 9-month period in sedentary overweight or obese adults. Body weight, systolic blood pressure (SBP) and diastolic blood pressure (DBP) were measured at baseline, month 4 and month 9. Both p-values from the Wald tests on replicated measurements of SBP and DBP are less than 0.001, which indicates that the variances of the measurement errors are not zero and there exit measurement errors in SBP and DBP. However, most popular methods for longitudinal data analysis require that data are measured perfectly, which in turn leads to naive approaches for handling mis-measured covariates in practice,

* Corresponding author at: Department of Biostatistics, School of Public Health and Key Laboratory of Public Health Safety, Fudan University, Shanghai, 200032, China.
E-mail address: gyqin@fudan.edu.cn (G. Qin).

https://doi.org/10.1016/j.csda.2017.11.010
0167-9473/

such as ignoring covariate measurement error or treating observed values of the covariates as if they were the true values. These naive approaches for handling mis-measured covariates would in general lead to biased inferences, even in the case of simple linear regressions (Buonaccorsi, 2010).

There have been discussions on how to treat covariate measurement error appropriately in the literature. For example, in a linear regression, one can correct the bias induced by covariate measurement error through correction for attenuation. In the context of semiparametric partially linear models, Liang et al. (1999) proposed a bias-corrected estimator that is a semiparametric version of the usual parametric correction for attenuation. For longitudinal data, many existing works adopted a structural approach for parameter estimation, specifying the distribution of the error-prone covariates. Carroll et al. (2012) considered mixed models with covariate measurement error and reviewed several maximum likelihood and pseudo-maximum likelihood estimating methods. See also Rummel et al. (2010) and Lin and Carroll (2006).

Furthermore, in the LEAN study, 21% of body mass index (BMI) measures, calculated from measured height and body weight, are missing, which raises more challenges to statistical estimations. Note, most methods mentioned above may not be easily tailored to dealing with such complex data. For example, the bias term induced by mis-measured covariates may not be easily calculated when there are outliers and dropouts at the same time. For the structural approach, it is a challenge to correctly specify the distribution of the error-prone covariates. Therefore, an estimation method that does not need the distribution specification of the mis-measured covariates is preferred. Assuming replicate measurements are available, Qin et al. (2016b) proposed a robust estimation method for longitudinal data with dropouts and measurement error, which can be easily implemented with popular softwares.

The main purpose of this paper is to develop a method with improved efficiency that can be easily implemented with standard algorithms developed for generalized estimating equations (GEE). Thus, it can be used widely in practice to handle complicated longitudinal data. Assuming that replicate measurements are available, we first propose an unbiased estimating equation and show that it is asymptotically more efficient than the estimator proposed in Qin et al. (2016a). Then we move forward to a more complicated scenario where missing responses, outliers and covariate measurement error exist simultaneously. The proposed method can be easily adapted to handle such scenarios, and we show that it is more efficient than the method in Qin et al. (2016b) through simulation studies.

The rest of the paper is organized as follows. Section 2 introduces models and methods. Section 2.1 includes the linear mean models and proposed unbiased estimating equation, which is asymptotically more efficient than the method in Qin et al. (2016a). In Section 2.2, the idea in the proposed method is extended to a more complicated case with simultaneous outliers, missing responses and covariate measurement error for partially linear marginal mean models. We illustrate through simulation studies that our method improves estimation efficiency compared to the method in Qin et al. (2016b) while retaining the consistency of the estimators when dealing with such complicated data. Simulation studies are conducted in Section 3. Real data analysis is given in Section 4, and we end with concluding remarks in Section 5. The details of regularity conditions and proofs are given in the online supplementary material.

2. Models and the proposed method

2.1. Linear model and measurement error process

Consider a study consisting of n subjects with m observations over time for each subject. Let Y_{ij} be the response and X_{ij} be the p-dimensional covariate vector for the ith subject at the jth observation. For simplicity, let $Y_i = (Y_{i1}, \ldots, Y_{im})^T$ be the response vector and $X_i = (X_{i1}, \ldots, X_{im})^T$ be the covariate matrix. We now consider the following linear marginal model:

$$Y_{ij} = X_{ij}^T \beta_0 + \epsilon_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, m, \tag{2.1}$$

where β_0 is a p-dimensional vector of unknown regression parameter, and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{im})^T$ is the vector of random error with mean zero and covariance matrix Σ_i.

In practice, the covariates X_{ij} may not be observed exactly. Instead of observing X_{ij}, we only observe W_{ij} as its surrogate. Here we assume the following classical additive error model for X_{ij} and W_{ij}:

$$W_{ij} = X_{ij} + \delta_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, m, \tag{2.2}$$

where δ_{ij} is the measurement error with mean zero and covariance matrix V, and δ_{ij} is independent of X_i and ϵ_i. Sometimes, repeated measurements are taken for the covariates in order to improve their accuracy. For simplicity, we assume that there are two replicate measurements, $W_{ij(1)}$ and $W_{ij(2)}$, for X_{ij}:

$$W_{ij(1)} = X_{ij} + \delta_{ij(1)}, W_{ij(2)} = X_{ij} + \delta_{ij(2)}, \tag{2.3}$$

where the measurement error $\delta_{ij(1)}$ and $\delta_{ij(2)}$ are assumed to be independent. For convenience, in the future we denote $W_{i(k)} = (W_{i1(k)}, \ldots, W_{im(k)})^T$, $\delta_{i(k)} = (\delta_{i1(k)}, \ldots, \delta_{im(k)})^T$ for $k = 1, 2$. If there are multiple repeated measurements, we can separate these measurements into two groups and use the average of the two groups as $W_{ij(1)}$ and $W_{ij(2)}$. In this way, we can reduce the measurement errors by average of the observations in each group and maintain the proposed asymptotic unbiased estimating equations using the independence among the measurement errors. In the case of multiple repeated measurements (e.g., K measurements), different separations will affect the estimation efficiency. Some simulation studies show that randomly separating two groups with nearly equal sizes, the integer part of $K/2$, will be a better choice which can result in more efficient estimation. Further work on this topic can be explored in the future.