Robust estimation in linear regression models for longitudinal data with covariate measurement errors and outliers

Yuexia Zhang, Guoyou Qin, Zhongyi Zhu, Jiajia Zhang

PII: S0047-259X(17)30683-8
DOI: https://doi.org/10.1016/j.jmva.2018.07.015
Reference: YJMVA 4395

To appear in: *Journal of Multivariate Analysis*

Received date : 17 November 2017

Please cite this article as: Y. Zhang, G. Qin, Z. Zhu, J. Zhang, Robust estimation in linear regression models for longitudinal data with covariate measurement errors and outliers, *Journal of Multivariate Analysis* (2018), https://doi.org/10.1016/j.jmva.2018.07.015

Robust estimation in linear regression models for longitudinal data with covariate measurement errors and outliers

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

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

Abstract

Measurement errors and outliers commonly arise during the process of longitudinal data collection and ignoring them in data analysis can lead to large deviations in estimates. Therefore, it is important to take into account the effect of measurement errors and outliers in longitudinal data analysis. In this paper, a robust estimating equation method for analyzing longitudinal data with covariate measurement errors and outliers is proposed. Specifically, the biases caused by measurement errors are reduced via using the independence between replicate measurements and the biases caused by outliers are corrected via centralizing the observed covariate matrix. The proposed method does not require specifying the distributions of the true covariates, response and measurement errors. In practice, it can be easily implemented via the standard generalized estimating equations algorithms. The asymptotic normality of the proposed estimator is established under regularity conditions. Extensive simulation studies show that the proposed method performs better in handling measurement errors and outliers than several existing methods. For illustration, the proposed method is applied to a data set from the Lifestyle Education for Activity and Nutrition (LEAN) study.

Keywords: Estimating equation, Longitudinal data, Measurement error, Outlier

1. Introduction

Longitudinal data are routinely collected in fields such as epidemiology and health research, and they are particularly important in investigating the longitudinal association between the outcome of interest and their potential risk factors. There are a variety of modeling and inference approaches for longitudinal data analysis; see [4, 6] and references therein. They perform well when the data are accurately collected. However, measurement errors and outliers are commonly generated during the process of longitudinal data collection. It is known that measurement errors can lead to inconsistent parameter estimation and invalid statistical inference. The same issues exist for outliers. The most popular longitudinal data analysis approaches, such as the likelihood method [9] and the generalized estimating equation (GEE) method [11], are very sensitive to outliers [2, 13]. Therefore, it is necessary to extend the standard longitudinal data analysis approaches by taking the effect of measurement errors or/and outliers into account.

When longitudinal data are subject to measurement errors only, many strategies can be used to rectify the problem. These strategies can be classified broadly into three categories: (i) methods which directly correct the biases of naive estimators; (ii) likelihood-based correction methods; and (iii) methods based on unbiased estimating functions. The first category includes the naive estimator correction method [21] and the simulation-extrapolation (SIMEX) method [25]; for the former it is sometimes difficult to gain the closed form of asymptotic bias while for the latter, it may be hard to choose the required specific extrapolation function. The likelihood-based correction method [10, 27] usually requires distributional assumptions for the true covariates, response and measurement errors; it is also sensitive to model misspecification. Difficulties in choosing appropriate distributions in practice limit its application

*Corresponding author
Email address: zhuzy@fudan.edu.cn (Zhongyi Zhu)