Journal of Multivariate Analysis 130 (2014) 409–424

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Variable selection and estimation for longitudinal survey data

Li Wang [a,*], Suojin Wang [b], Guannan Wang [a]

[a] Department of Statistics, University of Georgia, Athens, GA 30602, United States

[b] Department of Statistics, Texas A&M University, College Station, TX 77843, United States

HIGHLIGHTS

- We develop a general strategy for model selection in longitudinal surveys.
- We propose a survey weighted penalized GEE to select significant variables.
- We apply the EF-bootstrap method to obtain standard errors for complex surveys.
- We find that survey weights should be accounted for informative sampling designs.

ARTICLE INFO

Article history:
Received 10 October 2012
Available online 20 May 2014

AMS 2000 subject classifications:
primary 62G08

Keywords:
Bootstrap
Generalized estimating equations
Penalty
Superpopulation
Sampling weights

ABSTRACT

There is wide interest in studying longitudinal surveys where sample subjects are observed successively over time. Longitudinal surveys have been used in many areas today, for example, in the health and social sciences, to explore relationships or to identify significant variables in regression settings. This paper develops a general strategy for the model selection problem in longitudinal sample surveys. A survey weighted penalized estimating equation approach is proposed to select significant variables and estimate the coefficients simultaneously. The proposed estimators are design consistent and perform as well as the oracle procedure when the correct submodel was known. The estimating function bootstrap is applied to obtain the standard errors of the estimated parameters with good accuracy. A fast and efficient variable selection algorithm is developed to identify significant variables for complex longitudinal survey data. Simulated examples are illustrated to show the usefulness of the proposed methodology under various model settings and sampling designs.

1. Introduction

In the past two decades, various longitudinal surveys have been undertaken, where sample subjects are observed successively over time. Some examples are the US National Compensation Survey, the International Price Program, the Survey of Income and Program Participation, the US Longitudinal Studies of Aging and a range of more specialized studies. These represent a very substantial investment in longitudinal resources, producing a diverse portfolio of research materials, and a vibrant national research culture that has a strong international visibility. Although many studies of these surveys focus on estimating means, totals, proportions or ratios for certain populations, longitudinal survey data are frequently used for the modeling and estimation of the relationship in regression analysis. For example, longitudinal social surveys are conducted in many countries to identify factors that have effects on unemployment status or income; many health surveys are aimed to gain insight of health determinants rather than estimating population totals or proportions.

* Corresponding author.
E-mail address: lilywang@uga.edu (L. Wang).

http://dx.doi.org/10.1016/j.jmva.2014.05.006
0047-259X/

Longitudinal surveys are usually stratified and often multistage with unequal probabilities of selection at certain stages. If some parts of the population are sampled more intensively than others and the survey sampling design is ignored in the model selection, statistical inferences drawn from the sample can be remarkably different from those drawn from the population.

In this paper, marginal models for longitudinal survey data are considered to tackle the design and longitudinal features simultaneously. Generalized estimating equations (GEE) proposed by [11] is a popular method for these models. [13] first introduced GEE for longitudinal survey data. [17] adapted the GEE approach of [13] to the analysis of ordinal longitudinal survey responses. [3] developed a pseudo-GEE approach for longitudinal surveys under a joint randomization framework and established the consistency of the resulting estimators.

In many surveys, a large number of auxiliary variables may be collected, and we may want to determine the "best" subset of the variables. Longitudinal survey data with a large number of covariates have become increasingly more common in many scientific disciplines. One representative example is the Canadian National Population Health Survey where the researchers are interested in linking common risk factors with the possibility of loss of independence among seniors. In this study, many variables, such as age, gender, smoking status, weight, height, chronic conditions, area of residence, etc., were measured over the years to describe the seniors' health status and lifestyles. In some other examples of longitudinal data, the number of variables measured on each individual or sampling unit may not be many, but if we consider various interaction effects, the number of predictors in the statistical model can still be large. In addition, knowing which variables are relevant gives insight into the nature of the survey design problem. For example, variable selection can be adopted to find stratification variables in the primary sampling unit selection process for many surveys.

Variable selection is an essential part of many statistical methods, yet has been less studied in survey sampling compared with other areas of applied statistics. This is partly due to the challenges created in joint consideration of the sampling scheme, multilevel correlation and variable selection. The problem of selecting auxiliary variables was considered by [16] in the model-assisted framework while [4,5] in the prediction framework. [21] proposed a Bayesian information criterion based method to select the auxiliary variables for use in the additive model-assisted framework. Although they are practically useful, these traditional selection procedures ignore stochastic errors inherited in the stages of variable selections [6]. The well studied shrinkage methods such as LASSO [18,19,7] are developed under non-survey settings and are inappropriate to select variables for data collected through complex sampling designs.

In this paper, we propose a consistent variable selection and estimation method for the marginal mean models for survey sampling based on the penalized estimating equation approach. To the best of our knowledge, this is the first attempt to consider this approach in sample surveys. We demonstrate that the proposed method performs as well as the oracle procedure that assumes the true submodel to be known.

The rest of the paper is organized as follows. Section 2 introduces the models for longitudinal survey data, discusses the main ideas of the penalized estimating equation approach, and provides the asymptotic properties of the penalized estimators. Section 3 discusses some implementation issues and provides the estimating function bootstrap variance estimators. The performance of the proposed variable selection method is studied via simulated data in Section 4. Section 5 summarizes the main results along with areas for future research. The proofs of the theorems along with technical lemmas are provided in the Appendix.

2. Methodology

2.1. General setting

Suppose that the finite population U_N consists of N individuals. In a longitudinal survey, a sample s of size n is selected at wave one using a specified sampling scheme, and observed over a specified number of time points. Let w_i be longitudinal weights attached to the ith sample. We assume that the w_is are first adjusted for unit nonresponse, then subjected to post stratification adjustment to ensure consistency.

Suppose that the ith respondent is observed for m_i occasions ($1 \leq m_i \leq m$). The data for the ith sample consist of $\left\{y_{ij}, \mathbf{x}_{ij}\right\}_{j=1}^{m_i}$, where y_{ij} is the response on occasion j, $\mathbf{x}_{ij} = \left(x_{ij1}, \ldots, x_{ijd}\right)^{\mathrm{T}}$ is a d-vector of covariates for each sampled individual. The marginal model assumes that the mean response $\mu_{ij} = E(y_{ij}|\mathbf{x}_{ij})$ is a function of $\mathbf{x}_{ij}$. In this paper, we assume that μ_{ij} depends on $\mathbf{x}_{ij}$ through a known monotonic and differentiable link function $g(\cdot)$, so that we get the generalized linear superpopulation model

$$\eta_{ij} \equiv g(\mu_{ij}) = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}, \tag{1}$$

which holds for the whole population, where $\boldsymbol{\beta}$ is a d-dimensional regression parameter. To avoid confusion, in the following let $\boldsymbol{\beta}_0 \equiv (\beta_{01}, \ldots, \beta_{0d})^{\mathrm{T}}$ be the true value of $\boldsymbol{\beta}$. Denote $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^{\mathrm{T}}, \boldsymbol{\beta}_{02}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_{01}$ is $d_a \times 1$ vector of the active superpopulation coefficients, and $\boldsymbol{\beta}_{02} \equiv \mathbf{0}$ is a $(d - d_a) \times 1$ vector of the inactive coefficients. Our main goal is to identify the d_a significant variables in model (1) and provide an accurate estimation for the non-zero coefficients. Our estimated d_a is the number of the remaining non-zero β_js from the iterative algorithm presented in Section 3.1.

The GEE approach is a class of estimating equations which take into account the correlation arising due to a longitudinal study design, resulting in the increased efficiency of standard error estimates. For simplicity, denote the sampled response