Journal of Multivariate Analysis 102 (2011) 723–740

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Empirical likelihood for semiparametric regression model with missing response data

Liugen Xue *, Dong Xue

College of Applied Sciences, Beijing University of Technology, Beijing 100124, China

ARTICLE INFO

Article history:
Received 3 November 2009
Available online 24 December 2010

AMS 2000 subject classification:
primary 62G05
secondary 62G15

Keywords:
Confidence interval
Empirical likelihood
Missing response data
Regression coefficient
Semiparametric regression model

ABSTRACT

A bias-corrected technique for constructing the empirical likelihood ratio is used to study a semiparametric regression model with missing response data. We are interested in inference for the regression coefficients, the baseline function and the response mean. A class of empirical likelihood ratio functions for the parameters of interest is defined so that undersmoothing for estimating the baseline function is avoided. The existing data-driven algorithm is also valid for selecting an optimal bandwidth. Our approach is to directly calibrate the empirical log-likelihood ratio so that the resulting ratio is asymptotically chi-squared. Also, a class of estimators for the parameters of interest is constructed, their asymptotic distributions are obtained, and consistent estimators of asymptotic bias and variance are provided. Our results can be used to construct confidence intervals and bands for the parameters of interest. A simulation study is undertaken to compare the empirical likelihood with the normal approximation-based method in terms of coverage accuracies and average lengths of confidence intervals. An example for an AIDS clinical trial data set is used for illustrating our methods.

1. Introduction

In clinical trials and observational studies, complete response data are often not available for every object. Missing response data may arise due to many circumstances, including treatment drop-out, study drop-out, mistimed measurements, study subjects failing to report to a clinic for monthly evaluations, respondents refusing to answer certain items on a questionnaire, loss of information caused by uncontrollable factors, and so forth. In such circumstances, the usual inference procedures for complete data sets cannot be applied directly. The most common technique used by data analysis is to naively exclude subjects with missing data, then perform a regression analysis with the remaining data. This is called a complete-case analysis. Because subjects with any missing variables are excluded, it is well known that the complete-case analysis can give highly inefficient estimates [9]. To increase efficiency, one imputes a plausible value for each missing datum and then analyzes the results as if they were complete data. Commonly used imputation methods for missing response values include linear regression imputation [4,17,18], nonparametric regression imputation [2,19,1], ratio imputation [14], semiparametric partially linear regression imputation [16,6], among others.

Let (X, T, Y) be a random vector such that X is a $d \times 1$ vector on R^d, T ranges over a nondegenerate compact one-dimensional interval I, and Y is a response variable influenced by the factors (X, T). Without loss of generality, it can be assumed that I is the unit interval $[0, 1]$. In practice, some Y values in a sample of size n may be missing, but X and T are observed completely. That is, the data consists of the incomplete observations $\{(X_i, T_i, Y_i, \delta_i), 1 \leq i \leq n\}$ from (X, T, Y, δ), where all the X_i's and T_i's are observed, and $\delta_i = 0$ if Y_i is missing, $\delta_i = 1$ otherwise. Throughout this paper, we assume that

* Corresponding author.
E-mail address: lgxue@bjut.edu.cn (L. Xue).

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2010.11.001

Y is missing at random (MAR), that is, $P(\delta = 1|X, T, Y) = P(\delta = 1|X, T)$. Let $\{(X_i, T_i, Y_i, \delta_i), 1 \le i \le n\}$ be independent and identically distributed (i.i.d.) observations. We assume that the data set can be modeled as

$$Y_i = X_i^{\mathrm{T}}\beta + g(T_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where β is a $d \times 1$ vector of unknown regression coefficients, $g(\cdot)$ is an unspecified baseline function, the errors ε_i are assumed to be i.i.d. with $E(\varepsilon_i|X_i, T_i) = 0$ almost surely, and ε_i and δ_i are independent for $1 \le i \le n$. The model is also known as a semiparametric regression model since it combines both parametric and nonparametric components.

Model (1.1) for missing response data has been studied in existing literature. See, for example, [16,20,6,13]. They proposed a class of semiparametric estimators for the parameter β, as well as for the response mean θ. The resulting estimators were shown to be consistent and asymptotically normal under general assumptions. They also used the empirical likelihood approach proposed by Owen [11,12] to construct the confidence region and interval for β and θ, respectively. Liang et al. [6] proposed an empirical likelihood-based statistic for β, which is shown to have a chi-squared distribution asymptotically. Wang et al. [16] constructed an empirical likelihood ratio for θ. Their idea is to firstly impute the missing Y-values by the semiparametric regression imputation and then construct a complete data empirical likelihood ratio from an imputed data set as if they were i.i.d. observations. However, the imputed data are not i.i.d. because a plug-in estimator is used. As a consequence, the empirical log-likelihood ratio for θ under imputation is asymptotically distributed as a scaled chi-square variable. Therefore, the empirical log-likelihood ratio cannot be directly applied to make statistical inference on the mean θ. This motivates them to adjust the ratio such that the adjusted empirical log-likelihood ratio is asymptotically chi-squared. The adjustment is performed by multiplying an adjustable factor to get an adjusted ratio. There are two issues with this: one is that the unknown adjustment factor is difficult to estimate efficiently; the other is that the undersmoothing involved in the estimation creates a difficulty in selecting bandwidth. To solve above two issues, we use the bias-correction method to calibrate the empirical likelihood ratios, and the obtained empirical likelihood ratios obey Wilks' theorem. Our approaches differ from those of Wang et al. [16] and Liang et al. [6].

Generally, for semiparametric regression model, the empirical log-likelihood ratio for parameter of interest is asymptotically noncentral chi-squared. The reason is that there exists a bias in the empirical likelihood ratio function, because the plug-in estimator for nonparametric component is used. Thus, the empirical likelihood ratio needs to be modified by using the bias-correction technique. The basic idea is to expunge the bias from the empirical likelihood ratio function by modifying an auxiliary random vectors, and the modified empirical likelihood ratio has the asymptotic central chi-squared distribution. The details can be found in Sections 2.2 and 3.2 and Remark 3. The bias-correction method has been used in existing literature, see, for example, [22–27].

In this paper, we use the bias-correction technique to construct the empirical likelihood ratios for β, $g(t)$ and θ, and show that any of these empirical likelihood ratios is asymptotic chi-squared. We also construct a class of estimators for β, $g(t)$ and θ, and obtain their asymptotic distributions. These results can be directly used to construct the confidence intervals or regions for β, $g(t)$ and θ. We also give the confidence intervals for every component of β and the confidence bands of $g(t)$. The following two desired features are worth mentioning. The first is that, by using the bias-correction technique and the semiparametric regression imputation scheme in constructing empirical likelihood ratios and estimators, undersmoothing for estimating the baseline function is avoided, and the existing data-driven algorithm can be used to select an optimal bandwidth. This overcomes the difficulty in selecting bandwidth. The second is that our approach is to directly calibrate the empirical likelihood ratio so that the resulting empirical log-likelihood ratio is asymptotically chi-squared. The ratio does not need to be multiplied by an adjustment factor. This avoids estimating the unknown adjustment factor.

The rest of this paper is organized as follows. In Section 2, a class of empirical likelihood ratios for β is constructed, and their asymptotic properties are shown. In Section 3, we propose two calibrated methods for constructing empirical likelihood ratios for $g(t)$, and give their asymptotic results. We also give the confidence bands of $g(t)$. In Section 4, we construct a bias-correction empirical likelihood ratio for θ, and study the maximum empirical likelihood estimator of θ. Section 5 illustrates the finite-sample performances by conducting some simulation studies. Section 6 give a real data example. Section 7 is concluding remarks. Proofs of the main theorems are given in Appendix A.

2. Empirical likelihood for the regression coefficients

In this section, we propose some methods for constructing the empirical likelihood ratios and the estimators of β, and study their asymptotic behaviors.

2.1. Empirical likelihood with complete-case data

Pre-multiplying (1.1) by the observation indicator, we have

$$\delta_i Y_i = \delta_i X_i^{\mathrm{T}}\beta + \delta_i g(T_i) + \delta_i\varepsilon_i, \quad i = 1, \ldots, n,$$

and taking conditional expectations given T_i, we get

$$E(\delta_i Y_i|T_i = t) = E(\delta_i X_i^{\mathrm{T}}|T_i = t)\beta + E(\delta_i|T_i = t)g(t),$$