Computational Statistics and Data Analysis 79 (2014) 91–112

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Empirical likelihood inference in linear regression with nonignorable missing response

Cuizhen Niu [a], Xu Guo [b,c], Wangli Xu [a,*], Lixing Zhu [b]

[a] School of Statistics, Renmin University of China, Beijing, China

[b] Department of Mathematics, Hong Kong Baptist University, Hong Kong

[c] College of Economics and Management, Nanjing University of Aeronautics and Astronautics, China

ARTICLE INFO

Article history:
Received 24 September 2013
Received in revised form 25 February 2014
Accepted 6 May 2014
Available online 28 May 2014

Keywords:
Empirical likelihood inference
Nonignorable missing response
Linear regression
Inverse probability weighted
Imputed empirical likelihood

ABSTRACT

Parameter estimation for nonignorable nonresponse data is a challenging issue as the missing mechanism is unverified in practice and the parameters of response probabilities need to be estimated. This article aims at applying the empirical likelihood to construct the confidence intervals for the parameters of interest in linear regression models with nonignorable missing response data and the nonignorable missing mechanism is specified as an exponential tilting model. Three empirical likelihood ratio functions based on weighted empirical likelihood and imputed empirical likelihood are defined. It is proved that, except one that is chi-squared distributed, all the others are asymptotically weighted chi-squared distributed whenever the tilting parameter is either given or estimated. The asymptotic normality for the related parameter estimates is also investigated. Simulation studies are conducted to evaluate the finite sample performance of the proposed estimates in terms of coverage probabilities and average widths for the confidence intervals of parameters. A real data analysis is analyzed for illustration.

1. Introduction

Consider the classical linear regression model

$$y_i = x_i^\tau \beta + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where β is a $d \times 1$ vector of unknown regression parameter and ε_i's are independent and identically distributed (i.i.d.) random errors with conditional mean $E(\varepsilon|X) = 0$. Throughout this paper, we focus on the situation that some of the responses y_i in a sample of size n may be missing and all the covariates or auxiliary variables x_i's are observed completely. In this way, we obtain the following incomplete observations

$$(x_i, y_i, \delta_i), \quad i = 1, \ldots, n,$$

where δ_i is a missing indicator for ith individual and $\delta_i = 0$ if y_i is missing, otherwise $\delta_i = 1$. Often, the missing mechanism missing at random (MAR) is a common assumption for statistical analysis in the presence of missing data and is reasonable in many practical situations. Nevertheless, sometimes, there may be a concern that nonresponse is related to the value of the unobserved outcome variable y_i itself, even after controlling for x_i. The MAR mechanism then would become invalid. For

* Corresponding author.
E-mail address: wxu.stat@gmail.com (W. Xu).

http://dx.doi.org/10.1016/j.csda.2014.05.005
0167-9473/

example, in the surveys about income or the history of committing, the nonresponse rates tend to be related to the values of nonresponses. Instead of the classical assumption MAR for missing data, the present paper assumes that missing response data are either nonignorable or not missing at random (NMAR).

Nonignorable missing is a ubiquitous existing problem in various disciplines, for example, medical research, clinical trials and longitudinal studies and in recent decades, there has been a number of literatures for the analysis of nonignorable missing values. Based on the exponential tilting model for the response probability, Kim and Yu (2011) proposed a semiparametric estimation method of mean functions with nonignorable missing data and derived the $\sqrt{n}$-consistency when the tilting parameter is either given or estimated. Zhao et al. (2013) applied the empirical likelihood to the inference of mean functionals with nonignorable missing response data when the inverse probability weighted methods with and without auxiliary information are used, and the asymptotic properties are systematically investigated. In longitudinal study, the classical maximum likelihood (ML) method has been extensively applied to analyze longitudinal missing data. To avoid sensitivity of ordinary ML estimates to extreme observations or outliers, Sinha (2012) suggested a robust method in the framework of the maximum likelihood for analyzing incomplete longitudinal data with generalized linear mixed models. Beyond that, Imai (2009) introduced an identification strategy for average treatment effect under the nonignorable assumption to analyze randomized experiments with a nonignorable missing binary outcome. In a sensitivity analysis, Xie et al. (2011) relaxed the linearity assumption for response probability and provided a semiparametric approach of the generalized additive model for analyzing nonignorable missing data. Their approach can avoid fitting any complicated semiparametric joint selection model. Lee and Tang (2006) considered a nonlinear structural equation model with nonignorable missing covariates and ordered categorical data, where the missingness mechanism was specified a logistic regression model.

As to nonignorable missing data, the underlying assumptions are difficult to verify in practice and the results of relevant statistical inference may be sensitive to these assumptions. Under this circumstance, parameter estimation for nonignorable nonresponse data is a challenge. To the best of our knowledge, few references focus on the inference for parameter β in linear regression with nonignorable missing response. The present paper focuses on this issue.

The empirical likelihood approach for constructing confidence intervals in nonparametric setting was introduced by Owen (1988, 1990). Since then, there has been a rich body of literature about relevant statistical inference based on the empirical likelihood technique. The empirical likelihood method owns its broad usage and widely research to a number of important advantages. As mentioned in Hall and La Scala (1990), the empirical likelihood technique does not impose prior constraints on the shape of the region and it does not require the construction of a pivotal quantity, besides, the region is range preserving and transformation respecting. Moreover, they are of natural shape and orientation since the regions are obtained by contouring a log-likelihood ratio. After that, Owen (1991) applied the empirical likelihood to linear regression and demonstrated that the empirical log-likelihood ratio is asymptotically a χ^2 variable. As to the construction of the confidence interval, Zhu and Xue (2006) studied the empirical likelihood-based inference for the parameters in a partially linear single-index model and first presented a bias correction to eliminate non-negligible bias caused by nonparametric estimation so as to achieve the standard χ^2-limit of the empirical likelihood function. For the missing data, Xue (2009a) developed an empirical likelihood method to study the construction of confidence intervals and regions for the parameters of interest in linear regression models with missing response data. Besides, Xue (2009b) elaborated the construction of the confidence interval for response mean based on the bias-corrected empirical likelihood ratio, where the missing response was imputed by a kernel regression method. Qin et al. (2009) raised a unified empirical likelihood approach for the case with the number of estimating equations greater than the number of unknown parameters.

The rest of this article is organized as follows. In Section 2, we present the construction of confidence intervals. The asymptotic normality for the estimates of the parameters and the asymptotic properties for the proposed empirical likelihood functions are investigated in Section 3. Simulation studies and a real data analysis are conducted to evaluate the finite sample performance of the proposed estimates in Sections 4 and 5, respectively. The concluding discussions are included in Section 6. Proofs of the asymptotic results are relegated in the Appendix.

2. Empirical likelihood-based inference

In this section, we propose three methods for the confidence interval construction of parameters in the following.

2.1. Weighted empirical likelihood

For an incomplete dataset $\{(x_i, y_i, \delta_i),\ i = 1, \ldots, n\}$ with δ_i being the missing datum indicator with the response probability $p(x_i, y_i)$:

$$p(x_i, y_i) = P(\delta_i = 1|x_i, y_i).$$

To construct the empirical likelihood function, the following auxiliary random vector based on the inverse probability weighted method is introduced:

$$z_{i,W} := z_{i,W}(\beta) = \frac{\delta_i}{p(x_i, y_i)} x_i(y_i - x_i^\tau \beta). \tag{2.1}$$