Journal of Multivariate Analysis 115 (2013) 33–47

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

The cluster bootstrap consistency in generalized estimating equations

Guang Cheng [a,*], Zhuqing Yu [a], Jianhua Z. Huang [b]

[a] Purdue University, West Lafayette, IN 47907, United States
[b] Texas A&M University, College Station, TX 77843, United States

ARTICLE INFO

Article history:
Received 20 June 2011
Available online 14 September 2012

AMS subject classifications:
62F40
62F25
62F12

Keywords:
Bootstrap consistency
Clustered/longitudinal data
Exchangeably weighted cluster bootstrap
Generalized estimating equations
One-step bootstrap

ABSTRACT

The cluster bootstrap resamples clusters or subjects instead of individual observations in order to preserve the dependence within each cluster or subject. In this paper, we provide a theoretical justification of using the cluster bootstrap for the inferences of the generalized estimating equations (GEE) for clustered/longitudinal data. Under the general exchangeable bootstrap weights, we show that the cluster bootstrap yields a consistent approximation of the distribution of the regression estimate, and a consistent approximation of the confidence sets. We also show that a computationally more efficient one-step version of the cluster bootstrap provides asymptotically equivalent inference.

Published by Elsevier Inc.

1. Introduction

To analyze the clustered/longitudinal data, [10] introduced the Generalized Estimating Equations (GEE) approach to take into account of the correlation structure within each cluster/subject without specifying the joint distribution of observations from a cluster/subject. The sandwich variance estimator is widely used for GEE in the asymptotic inference of the regression parameters, e.g. construction of confidence sets, and is robust to mis-specification of the correlation structure. However, many empirical studies have shown that the sandwich estimator is usually downward biased and the bias could be substantial when the sample size is small, especially for binary responses [15,19,12]. To correct the underestimation by the sandwich estimator, modifications to the sandwich estimator have been investigated [12,9]. However, these bias corrections are only approximations and may be computationally unstable. Moreover, the asymptotic normality, which serves as the theoretical basis of using the sandwich estimator, need not be a good approximation when the number of subjects is small.

Therefore, resampling methods have been proposed in the literature to overcome the limitation of using the asymptotic normal inference and the sandwich estimator. Sherman and le Cessie [19] proposed the cluster bootstrap, which resamples subjects of a longitudinal data set, and argued that, by resampling subjects, the correlation structure within each subject is maintained and the bootstrap confidence intervals are produced in an automatic way so that the correlation structure can be left unspecified. Their simulation study showed that the bootstrap intervals are superior to the normal confidence interval built upon the sandwich estimator of variances. The superior empirical performances of bootstrapping longitudinal data were also reported in [14,7,3]. Despite various empirical evidences supporting the practical use of the cluster bootstrap, as far as we are aware, there is no theoretical study of this method.

In this paper, we provide a theoretical justification of using the cluster bootstrap for inference in GEE. We show that, under reasonable regularity conditions, the cluster bootstrap yields a consistent approximation of the distribution of the regression estimate, and a consistent approximation of the confidence sets. We establish our theoretical results under the

* Corresponding author.
E-mail addresses: chengg@stat.purdue.edu, chengg@purdue.edu (G. Cheng), zqyu@purdue.edu (Z. Yu), jianhua@stat.tamu.edu (J.Z. Huang).

0047-259X/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.jmva.2012.09.003

general setup of the exchangeably weighted cluster bootstrap, which contains the usual resample-cluster bootstrap as a special case. By choosing an appropriate weighting scheme, the general cluster bootstrap is useful to handle the difficult cases when the usual resample-cluster bootstrap breaks down, such as the zero or near-zero cell counts of the resamples for binary responses, when the number of subjects is small.

This paper also studies a computationally more efficient version of the cluster bootstrap. The cluster bootstrap is computationally expensive, because one needs to solve a new estimating equation for each bootstrap sample and the number of bootstrap samples is usually chosen to be quite large in order to achieve the desired inference accuracy. The one-step cluster bootstrap, which only computes one Gauss–Newton step for each bootstrap sample, is computationally more efficient because it avoids the full iteration in solving the estimating equations. We show that the one-step cluster bootstrap is asymptotically equivalent to the cluster bootstrap based on the full iteration. Our simulation results further provide empirical evidence.

The rest of the paper is organized as follows. Section 2 reviews the GEE formulation and gives rigorous asymptotic analysis on the regression estimate by extending the work by Xie and Yang [25]. Section 3 describes the cluster bootstrap sampling schemes and provides their theoretical justifications. Section 4 presents empirical results in terms of both simulated data and real data.

2. Generalized estimating equation

2.1. GEE formulation

Assume that the observations on different subjects are independent and the observations on the same subject are correlated. Let $Y_i = (y_{i1}, y_{i2}, \ldots, y_{im_i})'$ be an m_i-vector of responses and $X_i = (x_{i1}, x_{i2}, \ldots, x_{im_i})'$ be the corresponding $m_i \times p$ covariate matrix for $1 \le i \le n$. Suppose the marginal mean and covariance matrix of Y_i conditional on covariates are $\mu_i(\beta_0)$ and $\Sigma_i(\beta_0)$, where $\beta \in \mathcal{B} \subset \mathbb{R}^p$. An appealing feature of GEE is the incorporation of the "working covariance matrix" $V_i(\alpha, \beta)$ into the inferential process, which avoids specifying the possibly complicated correlation structure within each subject. $V_i(\alpha, \beta)$ is usually expressed as the form $A_i(\beta)R_i(\alpha)A_i(\beta)$, where $A_i^2(\beta_0)$ is a diagonal matrix of variance for Y_i and $R_i(\alpha)$ is the "working" correlation matrix fully specified by a nuisance vector $\alpha \in \mathcal{A} \subset \mathbb{R}^s$. The GEE introduced in [10] is of the following form:

$$U_n(\alpha, \beta) = \sum_{i=1}^{n} U_{ni}(\alpha, \beta) = \sum_{i=1}^{n} D_i'(\beta)V_i^{-1}(\alpha, \beta)S_i(\beta), \tag{1}$$

where $D_i(\beta) = \partial\mu_i(\beta)/\partial\beta$ and $S_i(\beta) = Y_i - \mu_i(\beta)$. Clearly, $U_{ni}(\alpha, \beta)$ is similar to the quasi-likelihood proposed in [24] except that the V_i is only a function of β. We say that $V_i(\alpha, \beta)$ is correctly specified if there exists a $\widetilde{\alpha} \in \mathcal{A}$ such that $V_i(\widetilde{\alpha}, \beta_0) = \Sigma_i(\beta_0)$, i.e., $R_i(\widetilde{\alpha})$ equals to the true correlation matrix R_{i0}, for any $i = 1, \ldots, n$. Here we give a concrete form of $U_{ni}(\alpha, \beta)$ when assuming the marginal distribution of y_{ij} conditional on the covariate x_{ij} follows the exponential family, i.e.

$$f(y_{ij}) = \exp\{y_{ij}\theta_{ij} - a(\theta_{ij}) + b(y_{ij})\}, \tag{2}$$

where $\theta_{ij} = h(x_{ij}'\beta)$ and h is a known injective function. In this case, we have

$$S_i(\beta) = Y_i - (\dot{a}(\theta_{i1}), \ldots, \dot{a}(\theta_{im_i}))', \tag{3}$$

$$A_i^2(\beta) = \mathrm{diag}(\ddot{a}(\theta_{ij})), \tag{4}$$

$$D_i(\beta) = A_i(\beta)\mathrm{diag}(\dot{h}(x_{ij}'\beta))X_i. \tag{5}$$

In this paper, we treat the correlation parameter α as the nuisance parameter. In practice, we can estimate α based on a preliminary estimate of β, e.g., $\widehat{\beta}_I$ under the working independence assumption, using the method of moment [10]. Wang and Carey [22,23] discussed the estimation of α for well-known correlation structures. Under mild conditions, we can easily obtain a root-n consistent $\widehat{\alpha}$. Therefore, we solve $\widehat{\beta}$ from the following estimated GEE:

$$U_n(\widehat{\alpha}, \beta) = 0, \tag{6}$$

where $\widehat{\alpha}$ is any root-n consistent estimate. The profile estimation approach employed in [10] is expected to improve only the second order efficiency of estimating β, and will also be discussed in this paper. We define α_0 as the limiting value of the estimator $\widehat{\alpha}$. If V_i is correctly specified, then $R_i(\alpha_0)$ is the true correlation matrix R_{i0} under regularity conditions. However, if V_i is not correctly specified, $R_i(\alpha_0)$ is unnecessarily R_{i0}.

2.2. Asymptotic results of GEE estimator

Liang and Zeger [10] proved the asymptotic normality of the GEE estimator $\widehat{\beta}$ under heuristic conditions. Xie and Yang [25] provided rigorous asymptotic analysis on the existence, consistency and asymptotic normality of $\widehat{\beta}$ that solves the