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

estimating equation $U_n(\alpha_0, \beta) = 0$. In this section, we consider the asymptotic behaviors of $\widehat{\beta}$ in the more realistic situation that α is unknown, and introduce necessary notations and technical tools for obtaining the bootstrap consistency results.

We denote the observations as $Z_i = (X_i, Y_i)$ for $i = 1, \ldots, n$, and view Z_i as the ith coordinate projection from the canonical probability space $(\mathcal{Z}^\infty, \mathcal{A}^\infty, P_Z^\infty)$ onto the ith copy of $\mathcal{Z}^\infty$. For simplicity, we assume X_i to be nonrandom. We write P_Z^∞ as P_Z for simplicity and write E_Z as the corresponding expectation. Let $\lambda_{\min}(T)(\lambda_{\max}(T))$ denote the smallest (largest) eigenvalue of the matrix T and $\mathcal{N}(\alpha_0) \times \mathcal{N}(\beta_0)$ be some neighborhood of (α_0, β_0). We define ordering between two square matrices as: $T_1 \geq T_2$ if and only if $\lambda' T_1 \lambda \geq \lambda' T_2 \lambda$ for any vector λ with $\|\lambda\| = 1$. The vector/matrix norm used in this paper is the Frobenius norm where $\|T\|^2 = \sum_i \sum_j T_{ij}^2$. The notations $\gtrsim$ and $\lesssim$ mean greater than, or smaller than, up to a universal constant.

We define the following notations

$$
\begin{aligned}
H_n(\alpha, \beta) &= \sum_{i=1}^{n} D_i'(\beta) V_i^{-1}(\alpha, \beta) D_i(\beta), \\
V_H(\alpha, \beta) &= \lim_{n\to\infty} \frac{H_n(\alpha, \beta)}{n}, \\
M_n(\alpha, \beta) &= \sum_{i=1}^{n} D_i'(\beta) V_i^{-1}(\alpha, \beta) \Sigma_i(\beta_0) V_i^{-1}(\alpha, \beta) D_i(\beta), \\
V_M(\alpha, \beta) &= \lim_{n\to\infty} \frac{M_n(\alpha, \beta)}{n}, \\
K_n(\alpha, \beta) &= -\frac{\partial U_n(\alpha, \beta)}{\partial \beta} = \sum_{i=1}^{n} K_{ni}(\alpha, \beta).
\end{aligned}
$$

For notational simplicity, we denote $H_n(\alpha_0, \beta_0)$ and $H_n(\widehat{\alpha}, \beta)$ as H_n and $H_n(\beta)$, respectively. The same rule applies to other notations, e.g. $V_i(\alpha, \beta)$ and $D_i(\beta)$.

We assume that $\min_{1\leq i\leq n} \lambda_{\min}(R_i(\alpha))$ is bounded away from zero for $\alpha \in \mathcal{N}(\alpha_0)$, and define a shrinking neighborhood of β_0 as follows

$$B_n(r) = \{\beta : \|H_n^{1/2}(\beta - \beta_0)\| \leq (m\tau_n)^{1/2} r\},$$

where $m = \max_{1\leq i\leq n} m_i$ independent of n and $\tau_n = \max_{1\leq i\leq n} \lambda_{\max}(R_i^{-1}(\alpha_0))$. We assume m to be bounded. In addition, we need the following conditions.

R1. $\lambda_{\min}(H_n)/\tau_n \to \infty$.

R2. *There exists a constant $c_0 > 0$ such that*

$$P_Z(K_n(\beta) \geq c_0 H_n) \longrightarrow 1$$

for all $\beta \in B_n(r)$ and any $r > 0$.

R3. *For any given $r > 0$,*

$$\sup_{\beta\in B_n(r)} \|H_n^{-1/2} K_n(\beta) H_n^{-1/2} - I\| = o_{P_Z}(1), \tag{7}$$

where I is a $p \times p$ identity matrix.

R4. *Let $y_i^\dagger = (y_{i1}^\dagger, \ldots, y_{im_i}^\dagger)' = A_i^{-1}(\beta_0)(y_i - \mu_i(\beta_0))$. There exists an $\epsilon > 0$, such that $E_Z(y_{ij}^\dagger)^{2+2/\epsilon}$ is uniformly bounded above, and*

$$(c_n \tau_n)^{1+\epsilon} \gamma_n \longrightarrow 0, \tag{8}$$

where $c_n = \lambda_{\max}(M_n^{-1} H_n)$ and $\gamma_n = \max_{1\leq i\leq n} \lambda_{\max}(H_n^{-1/2} D_i' V_i^{-1} D_i H_n^{-1/2})$.

R5. $\widehat{\alpha}$ is root-n consistent, i.e., $\sqrt{n}(\widehat{\alpha} - \alpha_0) = O_{P_Z}(1)$.

Conditions R1 and R4 are exactly Conditions (I_w^*) and (N_δ) in [25]. By considering the identity that $B^{-1} - A^{-1} = -A^{-1}(B - A)B^{-1}$ and the assumption that $\min_{1\leq i\leq n} \lambda_{\min}(R_i(\alpha))$ is bounded away from zero for $\alpha \in \mathcal{N}(\alpha_0)$, we have

$$\|R_i^{-1}(\alpha) - R_i^{-1}(\alpha_0)\| \lesssim \|R_i(\alpha) - R_i(\alpha_0)\|$$

for $\alpha \in \mathcal{N}(\alpha_0)$. Assuming that each element of $R_i(\alpha)$ is continuous around α_0 (or, more formally, $\max_{1\leq i\leq n} \|R_i(\alpha) - R_i(\alpha_0)\| \to 0$ if $\|\alpha - \alpha_0\| \to 0$), the following Theorem 2.1 is still valid after we replace $K_n(\beta)$ in Conditions R2–R3 by $K_n(\alpha_0, \beta)$ due to the consistency of $\widehat{\alpha}$. After this replacement, Conditions R2–R3 thus become Conditions (L_w^*) and (CC) in [25]. Therefore, we refer the detailed discussions on Conditions R1–R4 to [25], i.e., Appendix C. In the end, we want to point out the connection between $K_n(\beta)$ and $H_n(\beta)$ to illustrate the rationale behind Condition R3. We decompose $K_n(\beta)/n - H_n(\beta)/n$ as

$$[E_Z K_n(\beta)/n - H_n(\beta)/n] + [K_n(\beta)/n - E_Z K_n(\beta)/n] = I_n(\beta) + II_n(\beta), \tag{9}$$

where $I_n(\beta)$ and $II_n(\beta)$ are of complicated forms and given in Appendix A of [25]. Note that the uniform law of large number theory, i.e., Theorem 8.2 in [16], implies

$$\sup_{(\alpha,\beta)\in\mathcal{N}(\alpha_0)\times\mathcal{N}(\beta_0)}\left|\frac{1}{n}\sum_{i=1}^{n}[K_{ni}(\alpha,\beta)-E_Z(K_{ni}(\alpha,\beta))]\right|=o_{P_Z}(1). \tag{10}$$

Applying (10) to $II_n(\beta)$, we obtain that $\sup_{\beta\in B_n(r)}|II_n(\beta)|\xrightarrow{P_Z}0$ since $\widehat{\alpha}$ is $\sqrt{n}$-consistent and $B_n(r)$ is a shrinking neighborhood of β_0. Thus, we expect Condition R3 to hold by assuming the continuity of $I_n(\beta)$ and noticing the fact that $I_n(\beta_0)=0$.

The following theorem states the existence, consistence and asymptotic normality of $\widehat{\beta}$ solved from (6) by extending the results in [25] which assume α to be known. We skip the proof of Theorem 2.1 since it is a special case of our bootstrap Theorem 3.1 by setting the bootstrap weight $W_{ni}=1$ a.s., see (16).

Theorem 2.1. *Suppose Conditions* R1, R2 *and* R5 *hold. If V_H and V_M are both nonsingular, there exists a sequence of consistent $\widehat{\beta}$, i.e.*

$$P_Z(U_n(\widehat{\alpha},\widehat{\beta})=0)\longrightarrow 1, \tag{11}$$

$$\widehat{\beta}\xrightarrow{P_Z}\beta_0. \tag{12}$$

If, further, Conditions R3–R4 are satisfied, then

$$M_n^{-1/2}H_n(\widehat{\beta}-\beta_0)=(M_n/n)^{-1/2}\mathbb{G}_nU+o_{P_Z}(1), \tag{13}$$

where $\mathbb{G}_nU=(1/\sqrt{n})\sum_{i=1}^{n}U_{ni}(\alpha_0,\beta_0)$. Consequently,

$$\sqrt{n}(\widehat{\beta}-\beta_0)\xrightarrow{d}N(0,V_G), \tag{14}$$

where $V_G=V_H^{-1}V_MV_H^{-1}$.

Obviously, V_G will simplify to the model based version $V_0\equiv V_H^{-1}$ if V_i is modeled correctly. The asymptotic covariance V_G is usually estimated by the so-called "sandwich" estimator $\widehat{V}_H^{-1}\widehat{V}_M\widehat{V}_H^{-1}$, where V_H and V_M are estimated by replacing α, β and $\Sigma_i(\beta_0)$ with $\widehat{\alpha}$, $\widehat{\beta}$ and $\widehat{\Sigma}_i=S_i(\widehat{\beta})S_i'(\widehat{\beta})$. In addition, we may obtain a more efficient estimate for β from the nonparametrically estimated GEE in which V_i is replaced by $\widehat{\Sigma}_i$. Balan and Schiopu-Kratina [1] have recently provided rigorous asymptotic analysis for this two-step procedure.

3. Bootstrap consistency

3.1. Cluster bootstrap sampling scheme

In this section, we introduce the *exchangeably weighted cluster bootstrap sampling*. We start from a special case, i.e., nonparametric bootstrap. Denote $(Z_1^*,\ldots,Z_n^*)$ as independent draws with replacement from the original cluster data. Define

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

where D_i^*, V_i^* and S_i^* are computed based on $\{Z_1^*,\ldots,Z_n^*\}$. We define the bootstrap estimator $\widehat{\beta}^*$ as the solution of

$$U_n^*(\widehat{\alpha}^*,\beta)=0. \tag{15}$$

The estimator $\widehat{\alpha}^*$ may be computed based on either the original observations or bootstrap sample. This nonparametric cluster bootstrap was considered in [19]. However, a drawback of the nonparametric resampling when the response is binary is that the method can break down when the number of subjects is small, due to zero or near-zero cell counts caused by resampling; see [12]. This motivates us to consider the general class of *exchangeably weighted bootstrap* that includes nonparametric bootstrap and its smooth alternative, i.e., Bayesian bootstrap, as special cases. This general resampling scheme was first proposed in [18], and extensively studied by Barbe and Bertail [2], who suggested the name "weighted bootstrap", and in [13,17]. The practical usefulness of the more general scheme is well-documented in the literature. For example, we may apply the Bayesian bootstrap when the nonparametric bootstrap breaks down in the above scenario. Chatterjee and Bose [4] considered other variations of nonparametric bootstrap using the term "generalized bootstrap", and applied them to the generalized linear models. However, their results do not apply to the cluster data where there exists correlation within each cluster/subject.

In the exchangeably weighted cluster bootstrap, we can reformulate the $U_n^*(\alpha, \beta)$ as

$$U_n^*(\alpha, \beta) = \sum_{i=1}^{n} W_{ni} U_{ni}(\alpha, \beta) = \sum_{i=1}^{n} W_{ni} D_i'(\beta) V_i^{-1}(\alpha, \beta) S_i(\beta), \tag{16}$$

where $\{W_{n1}, \ldots, W_{nn}\}$ are the bootstrap weights. Now, we can formally define the bootstrap estimator $\widehat{\beta}^*$ as the solution of

$$U_n^*(\widehat{\alpha}^*, \beta) = 0, \tag{17}$$

where $\widehat{\alpha}^*$ may be computed based on either the original observations or the bootstrap sample. Note that the above nonparametric bootstrap estimator is a special case of the general formulation when $(W_{n1}, \ldots, W_{nn}) \sim$ Multinomial$(n, (n^{-1}, \ldots, n^{-1}))$. To obtain general results for the exchangeably weighted cluster bootstrap, we assume that the bootstrap weights $\{W_{ni}\}_{i=1}^n$ defined on the probability space $(\mathcal{W}, \Omega, P_W)$ satisfy the following Conditions W1–W5 [17]:

W1. *The vector* $W_n = (W_{n1}, \ldots, W_{nn})'$ *is exchangeable for all* $n = 1, 2, \ldots$, *i.e. for any permutation* $\pi = (\pi_1, \ldots, \pi_n)$ *of* $(1, 2, \ldots, n)$, *the joint distribution of*

$$\pi(W_n) = (W_{n\pi_1}, \ldots, W_{n\pi_n})'$$

is the same as that of W_n.

W2. $W_{ni} \geq 0$ *for all* n, i *and* $\sum_{i=1}^n W_{ni} = n$ *for all* n.

W3. $\limsup_{n\to\infty} \|W_{n1}\|_{2,1} \leq C < \infty$, *where* $\|W_{n1}\|_{2,1} = \int_0^\infty \sqrt{P_W(W_{n1} \geq u)}du$.

W4. $\lim_{\lambda\to\infty} \limsup_{n\to\infty} \sup_{t\geq\lambda} t^2 P_W(W_{n1} > t) = 0$.

W5. $(1/n)\sum_{i=1}^n (W_{ni} - 1)^2 \xrightarrow{P_W} c^2 > 0$.

The bootstrap weights corresponding to nonparametric bootstrap satisfy W1–W5. Another important class of bootstrap whose weights satisfy W1–W5 is the *multiplier bootstrap* in which $W_{ni} = \omega_i/\bar{\omega}_n$ and $(\omega_1, \ldots, \omega_n)$ are i.i.d. positive r.v.s with $\|\omega_1\|_{2,1} < \infty$. By taking $\omega_i \overset{i.i.d.}{\sim}$ Exp(1), we obtain the *Bayesian bootstrap* of [18]. The multiplier bootstrap is often thought to be a smooth alternative to the nonparametric bootstrap. In general, Conditions W3–W5 are easily satisfied under some lower moment conditions on W_{ni}; see Lemma 3.1 of [17]. The sampling schemes that satisfy Conditions W1–W5 include *the double bootstrap, the urn bootstrap and the grouped or delete-h Jackknife*; see [17]. By the fact that $(1/2)\|\cdot\|_2 \leq \|\cdot\|_{2,1}$, Condition W3 implies

$$\limsup_{n\to\infty} \|W_{n1}\|_2 \leq 2C < \infty. \tag{18}$$

The value of c in W5 is independent of n and depends on the resampling method, e.g., $c = 1$ for the nonparametric bootstrap and Bayesian bootstrap, and $c = \sqrt{2}$ for the double bootstrap. For the clustered data, the total number of measurements over all the subjects is $\sum_{i=1}^n m_i$. It is worth pointing out that the expected total number of measurements in the exchangeably weighted bootstrap, i.e., $E(\sum_{i=1}^n W_{ni} m_i)$, is also $\sum_{i=1}^n m_i$ since $EW_{ni} = 1$.

Remark 3.1. Li and Wang [11] proposed a smooth bootstrap idea in which they solve a perturbed estimating equation. Different from our bootstrap approach, their method does not correspond to any real resampling algorithm used in practice.

3.2. Bootstrap consistency

In this section, we present the main contribution of this paper. Specifically, as our first main result, we show that the exchangeably weighted bootstrap distribution of $(\sqrt{n}/c)(\widehat{\beta}^* - \widehat{\beta})$, conditional on the observed data, asymptotically imitates the distribution of $\sqrt{n}(\widehat{\beta} - \beta_0)$. As a consequence, we also establish the consistency of the bootstrap confidence set of β where consistency means that the coverage probability converges to the nominal level. In practice, $\widehat{\beta}$ or $\widehat{\beta}^*$ is usually obtained by the Gauss–Newton method after a few iterations. As our second main result, we will show that one-step iteration is actually sufficient to achieve the same efficiency as the fully iterative estimator. This result may be viewed as the bootstrap version of the one-step MLE result in the literature; see [8]. Surprisingly, in addition to the conditions for the asymptotic normality, we only need one more mild condition, i.e. Condition (22), to guarantee the consistency of cluster bootstrap. Thus, we can claim that the theoretical validity of cluster bootstrap is almost automatically guaranteed once the GEE estimator $\widehat{\beta}$ is shown to be asymptotically normal.

To rigorously state our results, we need to clarify that there exist two sources of randomness for the bootstrapped quantity $\widehat{\beta}^*$: one comes from the observed data; another comes from the resampling done by the bootstrap, i.e., randomness in W_{ni}'s. For the joint randomness involved, the product probability space is defined as

$$(\mathcal{Z}^\infty, \mathcal{A}^\infty, P_Z) \times (\mathcal{W}, \Omega, P_W) = (\mathcal{Z}^\infty \times \mathcal{W}, \mathcal{A}^\infty \times \Omega, P_{ZW}).$$

In this paper, we assume that the bootstrap weights W_{ni}'s are independent of the data Z_i's, thus $P_{ZW} = P_Z \times P_W$. Define E_{ZW} as the expectation w.r.t. P_{ZW}. The notations $E_{W|Z}$, E_Z and E_W are defined similarly. Given a real-valued function Δ_n defined

on the above product probability space, we say that Δ_n is of an order $o_{P_W}(1)$ in P_Z-probability if for any $\epsilon, \delta > 0$,

$$P_Z\{P_{W|Z}(|\Delta_n| > \epsilon) > \delta\} \longrightarrow 0 \quad \text{as } n \to \infty, \tag{19}$$

and that Δ_n is of an order $O_{P_W}(1)$ in P_Z-probability if for any $\delta > 0$, there exists a $0 < M < \infty$ such that

$$P_Z\{P_{W|Z}(|\Delta_n| \geq M) > \delta\} \longrightarrow 0 \quad \text{as } n \to \infty. \tag{20}$$

Given a function Γ_n defined only on $(\mathcal{Z}^\infty, \mathcal{A}^\infty, P_Z^\infty)$, if it is of an order $o_{P_Z}(1)(O_{P_Z}(1))$, then it is also of an order $o_{P_{ZW}}(1)(O_{P_{ZW}}(1))$ based on the following argument:

$$\begin{aligned} P_{ZW}(|\Gamma_n| > \epsilon) &= E_{ZW}1\{|\Gamma_n| > \epsilon\} = E_Z E_{W|Z}1\{|\Gamma_n| > \epsilon\} \\ &= E_Z 1\{|\Gamma_n| > \epsilon\} = P_Z\{|\Gamma_n| > \epsilon\}, \end{aligned}$$

where the third equation holds since Γ_n does not depend on the bootstrap weight. More results on transition of various stochastic orders are summarized in Lemma A.1 of the Appendix. Such results are used repeatedly in proving our bootstrap consistency theorem. To establish the bootstrap consistency, we need some measurability conditions such that Fubini's theorem can be used freely. However, there are enough assumptions in our model to guarantee this measurability. We omit further discussion of the measurability issue in this paper and refer interested readers to [5].

Define $K_n^*(\alpha, \beta)$ and $H_n^*(\alpha, \beta)$ as the bootstrap counterpart to $K_n(\alpha, \beta)$ and $H_n(\alpha, \beta)$, respectively. For simplicity, we write $K_n^*(\alpha_0, \beta_0)$ and $K_n^*(\widehat{\alpha}^*, \beta)$ as K_n^* and $K_n^*(\beta)$, respectively. The same rule also applies to $U_n^*(\alpha, \beta)$ and $H_n^*(\alpha, \beta)$.

We need the following bootstrap consistency Conditions R2′, R3′ and R5′, which are parallel to R2, R3 and R5.

R2′. *There exists a constant $c_0 > 0$ such that*

$$P_{ZW}(K_n^*(\beta) \geq c_0 H_n) \longrightarrow 1$$

for all $\beta \in B_n(r)$ and any $r > 0$.

R3′. *For any given $r > 0$, we assume*

$$\sup_{\beta \in B_n(r)} \|H_n^{-1/2} K_n^*(\beta) H_n^{-1/2} - I\| = o_{P_{ZW}}(1). \tag{21}$$

R5′. *$\widehat{\alpha}^*$ is root-n consistent, i.e., $\sqrt{n}(\widehat{\alpha}^* - \alpha_0) = O_{P_{ZW}}(1)$.*

Conditions R2′–R3′ can be implied by R2–R3 given the following simple condition, i.e.,

$$\sup_{\beta \in B_n(r)} \left| \frac{K_n^*(\beta) - K_n(\beta)}{n} \right| = o_{P_{ZW}}(1). \tag{22}$$

We can easily verify (22) by considering (10) and the following two equations:

$$\sup_{(\alpha,\beta) \in \mathcal{N}(\alpha_0) \times \mathcal{N}(\beta_0)} \left| \frac{1}{n} \sum_{i=1}^n [W_{ni} (K_{ni}(\alpha, \beta) - E_Z(K_{ni}(\alpha, \beta)))] \right| = o_{P_W}(1) \quad \text{in } P_Z\text{-prob.}, \tag{23}$$

$$\sup_{(\alpha,\beta) \in \mathcal{N}(\alpha_0) \times \mathcal{N}(\beta_0)} \left| \frac{1}{n} \sum_{i=1}^n (W_{ni} - 1) E_Z K_{ni}(\alpha, \beta) \right| = o_{P_{ZW}}(1), \tag{24}$$

where (23) and (24) follow from Lemma 3.6.16 in [21] and Theorem 8.2 of [16], respectively. Note that in the above we also use the facts that $o_{P_Z}(1)$ is $o_{P_{ZW}}(1)$ and that $o_{P_W}(1)$ in P_Z-Probability is $o_{P_{ZW}}(1)$ based on Lemma A.1. Condition (R5′) poses no difficulty in practice and is met trivially by, e.g. any root-n consistent $\widehat{\alpha}$ used in (6).

Theorem 3.1. *Suppose Conditions* R1, R2′, R5′ *and* W1–W5 *hold. If V_H and V_M are both nonsingular, there exists a sequence of consistent $\widehat{\beta}^*$, i.e.*

$$P_{ZW}(U_n^*(\widehat{\alpha}^*, \widehat{\beta}^*) = 0) \longrightarrow 1, \tag{25}$$

$$\widehat{\beta}^* \xrightarrow{P_{ZW}} \beta_0. \tag{26}$$

If, further, Conditions R3′ –R4′ *are satisfied, then*

$$M_n^{-1/2} H_n(\widehat{\beta}^* - \widehat{\beta}) = (M_n/n)^{-1/2} \mathbb{G}_n^* U + o_{P_W}(1) \quad \text{in } P_Z\text{-probability}, \tag{27}$$

where $\mathbb{G}_n^* U = (1/\sqrt{n}) \sum_{i=1}^n (W_{ni} - 1) U_{ni}(\alpha_0, \beta_0)$. Consequently,

$$\sup_{z \in \mathbb{R}^p} \left| P_{W|\mathcal{Z}_n}((\sqrt{n}/c)(\widehat{\beta}^* - \widehat{\beta}) \leq z) - P(N(0, V_G) \leq z) \right| \xrightarrow{P_Z} 0, \tag{28}$$

where $\leq$ is taken componentwise, c is given in W5. Thus, we have

$$\sup_{z \in \mathbb{R}^p} \left| P_{W|\mathcal{Z}_n}((\sqrt{n}/c)(\widehat{\beta}^* - \widehat{\beta}) \leq z) - P_Z(\sqrt{n}(\widehat{\beta} - \beta_0) \leq z) \right| \xrightarrow{P_Z} 0. \tag{29}$$

Remark 3.2. Our results in Theorems 2.1 and 3.1 also apply to the profile estimation of β proposed in [10]. Specifically, we first estimate α as a function of β, denoted as $\widehat{\alpha}(\beta)$, according to different parameterizations of $R_i(\alpha)$, and then estimate β by solving $U_n(\widehat{\alpha}(\beta), \beta) = 0$. We can easily adapt our proof of Theorem 3.1 to accommodate the profile version of the GEE by invoking the chain rule arguments under the following assumptions:

(a) $\widehat{\alpha}^*(\beta)$ is consistent for any $\beta \in B_n(r)$;
(b) $\widehat{\alpha}^*(\beta_0)$ is $\sqrt{n}$ consistent;
(c) $\partial\widehat{\alpha}^*(\beta)/\partial\beta = O_{P_{ZW}}(1)$ for any $\beta \in B_n(r)$.

These assumptions are very mild since $\widehat{\alpha}(\beta)$ (and thus $\widehat{\alpha}^*(\beta)$) is usually expressed as a smooth function of β; see Section 4 of [10].

The distribution consistency results proven in Theorem 3.1 imply the consistency of a variety of bootstrap confidence sets, i.e., *percentile*, *hybrid* and *t* types. A lower γ-th quantile of bootstrap distribution is any quantity $\tau^*_{n\gamma} \in \mathbb{R}^p$ satisfying $\tau^*_{n\gamma} = \inf\{\epsilon : P_{W|\mathbb{Z}_n}(\widehat{\beta}^* \le \epsilon) \ge \gamma\}$, where ϵ is an infimum over the given set only if there does not exist a $\epsilon_1 < \epsilon$ in $\mathbb{R}^p$ such that $P_{W|\mathbb{Z}_n}(\widehat{\beta}^* \le \epsilon_1) \ge \gamma$. Because of the assumed smoothness of $U_{ni}(\alpha, \beta)$ in our setting, we can, without loss of generality, assume $P_{W|\mathbb{Z}_n}(\widehat{\beta}^* \le \tau^*_{n\gamma}) = \gamma$. Due to (29), we can approximate the γ-th quantile of the distribution of $(\widehat{\beta} - \beta_0)$ by $(\tau^*_{n\gamma} - \widehat{\beta})/c$. Thus we define the *percentile*-type bootstrap confidence set as

$$BC_p(\gamma) = \left[\widehat{\beta} + \frac{\tau^*_{n(\gamma/2)} - \widehat{\beta}}{c}, \widehat{\beta} + \frac{\tau^*_{n(1-\gamma/2)} - \widehat{\beta}}{c}\right]. \tag{30}$$

Similarly, we can approximate the γ-th quantile of $\sqrt{n}(\widehat{\beta} - \beta_0)$ by $\kappa^*_{n\gamma}$, where $\kappa^*_{n\gamma}$ is the γ-th quantile of the hybrid quantity $(\sqrt{n}/c)(\widehat{\beta}^* - \widehat{\beta})$, i.e., $P_{W|\mathbb{Z}_n}((\sqrt{n}/c)(\widehat{\beta}^* - \widehat{\beta}) \le \kappa^*_{n\gamma}) = \gamma$. Thus we define the *hybrid*-type bootstrap confidence set as

$$BC_h(\gamma) = \left[\widehat{\beta} - \frac{\kappa^*_{n(1-\gamma/2)}}{\sqrt{n}}, \widehat{\beta} - \frac{\kappa^*_{n(\gamma/2)}}{\sqrt{n}}\right]. \tag{31}$$

Note that $\tau^*_{n\gamma}$ and $\kappa^*_{n\gamma}$ are not unique since β is assumed to be a vector.

We now prove the consistency of the above bootstrap confidence sets by using the arguments in Lemma 23.3 of [20]. First, it follows from (14) and (28) that, for any $z \in \mathbb{R}^p$,

$$P_Z(\sqrt{n}(\widehat{\beta} - \beta_0) \le z) \longrightarrow \Psi(z), \tag{32}$$

$$P_{W|\mathbb{Z}_n}((\sqrt{n}/c)(\widehat{\beta}^* - \widehat{\beta}) \le z) \xrightarrow{P_Z} \Psi(z), \tag{33}$$

where $\Psi(z) = P(N(0, V_G) \le z)$. The quantile convergence theorem, i.e., Lemma 21.2 in [20], applied to (33) implies that $\kappa^*_{n\gamma} \to \Psi^{-1}(\gamma)$ almost surely. When applying quantile convergence theorem, we use the almost sure representation Theorem 2.19 in [20] and argue along subsequences. Then, the Slutsky's Lemma implies that $\sqrt{n}(\widehat{\beta} - \beta_0) - \kappa^*_{n(\gamma/2)}$ weakly converges to $N(0, V_G) - \Psi^{-1}(\gamma/2)$. Thus,

$$\begin{aligned} P_{ZW}\left(\beta_0 \le \widehat{\beta} - \frac{\kappa^*_{n(\gamma/2)}}{\sqrt{n}}\right) &= P_{ZW}\left(\sqrt{n}(\widehat{\beta} - \beta_0) \ge \kappa^*_{n(\gamma/2)}\right) \\ &\to P_{ZW}\left(N(0, V_G) \ge \Psi^{-1}(\gamma/2)\right) = 1 - \gamma/2. \end{aligned}$$

This argument yields the consistency of the *hybrid*-type bootstrap confidence set, i.e., (35) below, and can also be applied to justify the *percentile*-type bootstrap confidence set, i.e., (34) below. The following Corollary 3.2 summarizes the above discussion.

Corollary 3.2. *Under the conditions in Theorem 3.1, we have*

$$P_{ZW}\left(\beta_0 \in BC_p(\gamma)\right) \longrightarrow 1 - \gamma, \tag{34}$$

$$P_{ZW}\left(\beta_0 \in BC_h(\gamma)\right) \longrightarrow 1 - \gamma, \tag{35}$$

as $n \to \infty$.

It is well known that the above bootstrap confidence sets can be computed easily through routine bootstrap sampling, and yield better empirical probability coverage in particular when the sample size is small.

Remark 3.3. Provided any consistent $\widehat{V}^*_G \xrightarrow{P_{ZW}} V_G$ and $\widehat{V}_G \xrightarrow{P_Z} V_G$, we can define the *t*-type bootstrap confidence set as

$$BC_t(\gamma) = \left[\widehat{\beta} - \frac{\widehat{V}_G^{1/2}\omega^*_{n(1-\gamma/2)}}{\sqrt{n}}, \widehat{\beta} - \frac{\widehat{V}_G^{1/2}\omega^*_{n(\gamma/2)}}{\sqrt{n}}\right],$$

where $\omega^*_{n\gamma}$ satisfies $P_{W|\mathcal{Z}_n}((\sqrt{n}/c)(\widehat{V}^*_G)^{-1/2}(\widehat{\beta}^* - \widehat{\beta}) \le \omega^*_{n\gamma}) = \gamma$. By applying again the arguments in Lemma 23.3 of [20], we can prove that

$$P_{ZW}(\beta_0 \in BC_t(\gamma)) \longrightarrow 1 - \gamma,$$

as $n \to \infty$.

In practice, $\widehat{\beta}^*$ is usually obtained by the Gauss–Newton method after a few iterations; see Section 3.2 of [10]. We will show that one-step iteration is actually sufficient to achieve the same theoretical efficiency as the fully iterative estimate. Let $\widehat{\beta}^*_0$ be any root-n consistent initial estimate, e.g., $\widehat{\beta}_I$. Denote $\widehat{\beta}^*_1$ as the one step cluster bootstrap estimator. Specifically, we have

$$\widehat{\beta}^*_1 = \widehat{\beta}^*_0 + [H^*_n(\widehat{\beta}^*_0)]^{-1} U^*_n(\widehat{\beta}^*_0). \tag{36}$$

Note that, for computational efficiency, we may fix $\widehat{\beta}^*_0$ in each bootstrap sample, and compute the H_n term only once by replacing $H^*_n(\widehat{\beta}^*_0)$ with $H_n(\widehat{\beta}^*_0)$ in (36). Such modifications will not affect the theoretical properties of $\widehat{\beta}^*_1$.

Our Theorem 3.3 below theoretically justifies the one-step bootstrap algorithm just described. Specifically, we show that $\widehat{\beta}^*_1$ is asymptotically equivalent to $\widehat{\beta}^*$, which implies one-step estimate share the same limiting distribution as the fully iterative estimate. This automatically enables us to build valid bootstrap confidence sets simply based on $\widehat{\beta}^*_1$.

Theorem 3.3. *Under Conditions in Theorem 3.1 and the Condition R6 that $V_H(\alpha, \beta)$ is continuous for $(\alpha, \beta) \in \mathcal{N}(\alpha_0) \times \mathcal{N}(\beta_0)$, we have*

1. $\|\widehat{\beta}^*_1 - \widehat{\beta}^*\| = o_{P_{ZW}}(n^{-1/2})$;
2. *The one-step percentile and hybrid type bootstrap confidence sets, in which $\widehat{\beta}^*$ is replaced by $\widehat{\beta}^*_1$ when calculating $\tau^*_{n\gamma}$ and $\kappa^*_{n\gamma}$ in (30) and (31), provide asymptotically correct probability coverage.*

Condition R6 is trivially satisfied since $\|D_i(\beta)\|$, $\|A_i(\beta)\|$ and $\|R_i(\alpha)\|$ are usually assumed to be smooth in practice. For example, we can assume that $\dot{h}(\cdot)$ and $\ddot{a}(\cdot)$ are smooth functions in the exponential family (2) to imply the above desired smoothness. In the end, we want to point out that Corollary 3.2 and Theorem 3.3 also hold under working independence assumption, i.e., $R_i(\alpha) = I$.

4. Empirical studies

4.1. Simulations

We conducted a simulation study to compare the coverage probabilities and the lengths of five different confidence intervals for β: (i) the GEE robust C.I. based on the asymptotic normality and the sandwich estimator of the variance; (ii) the one-step percentile bootstrap C.I.; (iii) the fully iterative percentile bootstrap C.I.; (iv) the one-step hybrid bootstrap C.I.; (v) the fully iterative hybrid bootstrap C.I.. Four simulation setups were considered.

In the first setup, data were generated from a regression model $Y_{ij} = \mu_{ij} + e_{ij}$, $1 \le i \le n$, $1 \le j \le m$, where i indicates the ith cluster and j indicates the jth observation within a cluster. The mean $\mu_{ij} = \exp(\beta_0 + \beta_1 x_{ij})$ with the true values $\beta_0 = 0$ and $\beta_1 = 1$, while the error term $e_{ij} = c_i + z_{ij}$ with $c_i \sim N(0, 1)$, $z_{ij} \sim N(0, 0.36)$, and c_i, z_{ij} being independent. The covariate x_{ij}'s are equidistantly distributed on $[-1, 1]$ for each $1 \le i \le n$. Under the above setup, the correlation structure is the exchangeable with $\mathrm{corr}(Y_{it}, Y_{it'}) = \alpha$ for all $t \ne t'$, and the true value $\alpha_0 = 25/34$. In our study, we considered two cases: (i) α was estimated as in [10]; and (ii) set $\widehat{\alpha} = 0$, corresponding to the working independence correlation structure.

In the second setup, Y_{ij}'s were drawn from a Poisson distribution with mean $\mu_{ij}\xi_{ij}$, where $\mu_{ij} = \exp(\beta_0 + \beta_1 x_{ij})$ with $(\beta_0, \beta_1)^T = (0, 1)^T$ and $\exp(x_{ij})$ uniformly distributed on $[j, j + 1]$, ξ_{ij} follows the Gamma distribution with mean 1 and variance 0.5. By allowing ξ_{ij} being correlated for $1 \le j \le m$, the covariance matrix of Y_i is $\mathrm{diag}(\{\mu_{ij}\}) + \mathrm{diag}(\{\mu_{ij}\})\mathrm{cov}(\xi_i)\,\mathrm{diag}(\{\mu_{ij}\})$. In our study, we let the correlation matrix of ξ_i be the one with all the off-diagonal terms 0.81. The correlated Gamma variables were then generated from $\xi_{ij} = \sum_{k=1}^{4} z^2_{kj}/4$ and each of the four m-vector (z_{kj}), $1 \le k \le 4$ was independently sampled from $N(0, R_0)$, where R_0 is a correlation matrix with all the off-diagonal terms being 0.9.

We considered logistic regression in the third setup where data were generated from the model: $\log p_{ij} = \log P\{Y_{ij} = 1\} = \exp(\beta_0 + \beta_1 x_{ij})$. Here Y_{ij}'s are binary random variables and $(\beta_0, \beta_1)^T = (0, 1)^T$. The covariates x_{ij}'s are uniformly distributed on $[-1, 1]$. We set the working correlation structure as exchangeable with the parameter $\alpha = 0.5$. As adopted in [19], to generate binary random variables with the specified correlation structure, the latent variables Z_{ij}'s were generated from a multivariate normal distribution with zero mean and the same correlation matrix. Then we let $Y_{ij} = 1$ if $Z_{ij} < \Phi^{-1}(p_{ij})$, where Φ denotes the c.d.f. of the standard normal distribution. The number of clusters n was taken to be 25, 30, 35 and the cluster size m was fixed as 5 in all the above setup examples.

To investigate the behavior of the cluster bootstrap when some of the assumptions of our theory are invalid, we considered $m = n$ in the fourth setup. Thus, the bounded cluster size condition is violated. Data were sampled from the regression model $Y_{ij} = \beta_0 + \beta_1 x_{ij} + e_{ij}$ with $\beta_0 = 0$ and $\beta_1 = 1$. For each i, the errors e_{ij}'s follows the autoregressive (AR) model of order 1 with the parameter $\alpha = 0.9$. The covariate x_{ij}'s are fixed and equidistant on $[-1, 1]$ for each $1 \le i \le n$. Similar to the previous three setups, the cluster size n was taken to be 25, 30, 35.

Table 1
The empirical coverage probabilities of various 90% C.I.s are shown together with the average interval length in parentheses. Computed based on 500 simulation runs.

Parameters	β_0			β_1		
Number of clusters	$n = 25$	$n = 30$	$n = 35$	$n = 25$	$n = 30$	$n = 35$
Simulation setup I: Gaussian regression model with log link. Cluster size $m = 5$. True working correlation = 'exchangeable'						
Working correlation = 'exchangeable'						
GEE robust	.876 (.628)	.854 (.510)	.854 (.537)	.860 (.436)	.848 (.452)	.856 (.491)
Bootstrap, percentile, one step iteration	.868 (.616)	.854 (.568)	.876 (.552)	.864 (.483)	.844 (.445)	.872 (.430)
Bootstrap, percentile, full iterations	.880 (.659)	.870 (.604)	.900 (.490)	.868 (.516)	.860 (.474)	.918 (.389)
Bootstrap, hybrid, one step iteration	.870 (.616)	.846 (.568)	.854 (.552)	.850 (.483)	.844 (.445)	.838 (.430)
Bootstrap, hybrid, full iterations	.850 (.659)	.852 (.604)	.880 (.490)	.838 (.516)	.838 (.474)	.884 (.389)
Working correlation = 'independence'						
GEE robust	.884 (.769)	.878 (.708)	.888 (.655)	.874 (.592)	.868 (.545)	.868 (.500)
Bootstrap, percentile, one step iteration	.876 (.713)	.870 (.658)	.864 (.611)	.866 (.548)	.862 (.506)	.842 (.466)
Bootstrap, percentile, full iterations	.868 (.722)	.878 (.665)	.876 (.617)	.858 (.556)	.874 (.513)	.874 (.472)
Bootstrap, hybrid, one step iteration	.858 (.713)	.886 (.658)	.886 (.611)	.842 (.548)	.866 (.506)	.862 (.466)
Bootstrap, hybrid, full iterations	.844 (.722)	.874 (.665)	.868 (.617)	.830 (.556)	.856 (.513)	.840 (.472)
Simulation setup II: Poisson regression with log link. Cluster size $m = 5$. True working correlation = 'exchangeable'						
Working correlation = 'exchangeable'						
GEE robust	.862 (.714)	.856 (.667)	.854 (.622)	.876 (.432)	.844 (.401)	.840 (.376)
Bootstrap, percentile, one step iteration	.884 (.729)	.856 (.664)	.864 (.615)	.892 (.441)	.876 (.401)	.884 (.373)
Bootstrap, percentile, full iterations	.888 (.741)	.854 (.672)	.870 (.622)	.888 (.448)	.886 (.406)	.890 (.378)
Bootstrap, hybrid, one step iteration	.896 (.729)	.850 (.664)	.864 (.615)	.890 (.441)	.882 (.401)	.886 (.373)
Bootstrap, hybrid, full iterations	.886 (.741)	.852 (.672)	.874 (.622)	.896 (.448)	.886 (.406)	.888 (.378)
Working correlation = 'independence'						
GEE robust	.856 (.714)	.848 (.666)	.872 (.624)	.866 (.438)	.854 (.404)	.874 (.381)
Bootstrap, percentile, one step iteration	.892 (.713)	.858 (.661)	.890 (.627)	.890 (.548)	.856 (.402)	.884 (.381)
Bootstrap, percentile, full iterations	.888 (.740)	.862 (.670)	.896 (.634)	.884 (.451)	.858 (.406)	.882 (.384)
Bootstrap, hybrid, one step iteration	.896 (.713)	.862 (.661)	.896 (.627)	.890 (.548)	.862 (.402)	.884 (.381)
Bootstrap, hybrid, full iterations	.896 (.740)	.866 (.670)	.898 (.634)	.894 (.451)	.860 (.406)	.888 (.384)
Simulation setup III: Logistic regression with logit link. Cluster size $m = 5$. True working correlation = 'exchangeable'						
Working correlation = 'exchangeable'						
GEE robust	.888 (.973)	.916 (.891)	.876 (.827)	.898(1.323)	.916(1.208)	.864(1.115)
Bootstrap, percentile, one step iteration	.890 (.925)	.906 (.840)	.896 (.799)	.882(1.121)	.892(1.024)	.884(0.887)
Bootstrap, percentile, full iterations	.876 (.959)	.896 (.864)	.902 (.780)	.856(1.068)	.886(0.969)	.882(0.948)
Bootstrap, hybrid, one step iteration	.876 (.925)	.898 (.840)	.920 (.799)	.882(1.121)	.868(1.024)	.892(0.887)
Bootstrap, hybrid, full iterations	.932 (.959)	.922 (.864)	.892 (.780)	.902(1.068)	.908(0.969)	.908(0.948)
Working correlation = 'independence'						
GEE robust	.872 (.970)	.912 (.889)	.876 (.825)	.898(1.396)	.898(1.280)	.866(1.179)
Bootstrap, percentile, one step iteration	.896 (.926)	.908 (.840)	.892 (.798)	.878(0.926)	.902(1.023)	.874(0.976)
Bootstrap, percentile, full iterations	.870 (.957)	.896 (.862)	.902 (.781)	.924(0.957)	.884(0.969)	.880(0.948)
Bootstrap, hybrid, one step iteration	.878 (.926)	.902 (.840)	.874 (.798)	.872(0.926)	.884(1.023)	.894(0.976)
Bootstrap, hybrid, full iterations	.924 (.957)	.884 (.862)	.880 (.781)	.902(0.957)	.908(0.969)	.892(0.948)
Simulation setup IV: Gaussian regression with identity link. Cluster size $m = n$. True working correlation = 'AR1'						
Working correlation = 'AR1'						
GEE robust	.884 (.058)	.878 (.044)	.892 (.035)	.884 (.162)	.878 (.125)	.892 (.100)
Bootstrap, percentile, one step iteration	.886 (.058)	.872 (.044)	.890 (.035)	.884 (.161)	.872 (.124)	.894 (.100)
Bootstrap, percentile, full iterations	.886 (.058)	.872 (.044)	.890 (.035)	.884 (.162)	.872 (.125)	.892 (.100)
Bootstrap, hybrid, one step iteration	.872 (.058)	.872 (.044)	.894 (.035)	.874 (.161)	.872 (.124)	.896 (.100)
Bootstrap, hybrid, full iterations	.876 (.058)	.876 (.044)	.894 (.035)	.878 (.162)	.880 (.125)	.896 (.100)
Working correlation = 'independence'						
GEE robust	.884 (.058)	.878 (.044)	.892 (.035)	.884 (.162)	.878 (.125)	.892 (.101)
Bootstrap, percentile, one step iteration	.886 (.059)	.872 (.044)	.890 (.035)	.884 (.162)	.872 (.125)	.894 (.100)
Bootstrap, percentile, full iterations	.886 (.059)	.872 (.044)	.890 (.035)	.884 (.162)	.872 (.125)	.894 (.100)
Bootstrap, hybrid, one step iteration	.878 (.059)	.876 (.044)	.894 (.035)	.878 (.162)	.880 (.125)	.896 (.100)
Bootstrap, hybrid, full iterations	.878 (.059)	.876 (.044)	.894 (.035)	.878 (.162)	.880 (.125)	.896 (.100)

For constructing the bootstrap C.I.s, we let the number of bootstrap resamples be $B = 1000$. The coverage probability was calculated as the proportion of the C.I.s (with nominal level 90%) covering β_0 and β_1 respectively out of 500 simulated data sets. The simulation results are summarized in Table 1. Under the Gaussian regression settings, the bootstrap C.I.s have comparable coverage performance with the robust C.I.s; this is expected for the Gaussian errors. For most cases of Poisson and logistic regression setups, the bootstrap C.I.s have coverage probabilities closer to the nominal 90% while

Table 2
Accumulated computing time (in seconds) for bootstrap confidence intervals in 500 simulation runs.

Number of clusters	$n = 25$	$n = 30$	$n = 35$
Simulation setup I: Gaussian regression model with log link Correlation = 'exchangeable'			
Full iterations	3796.219	4100.124	4463.620
One step iteration	2260.707	2285.726	2496.801
Correlation = 'independence'			
Full iterations	2436.162	2561.358	2749.852
One step iteration	2267.747	2336.769	2472.898
Simulation setup II: Poisson regression with log link Correlation = 'exchangeable'			
Full iterations	3374.469	3862.878	4118.248
One step iteration	2346.223	2620.854	2755.763
Correlation = 'independence'			
Full iterations	2436.162	2561.358	2788.320
One step iteration	2267.747	2336.769	2401.602
Simulation setup III: logistic regression with logit link Correlation = 'exchangeable'			
Full iterations	5179.339	5683.212	6215.173
One step iteration	3690.654	4057.560	4262.934
Correlation = 'independence'			
Full iterations	3590.903	3527.743	3690.815
One step iteration	3023.197	3175.955	3361.186
Simulation setup IV: Gaussian regression with identity link Correlation = 'exchangeable'			
Full iterations	3374.469	3862.878	4118.248
One step iteration	2346.223	3370.272	2755.763
Correlation = 'independence'			
Full iterations	3063.152	3237.342	3629.445
One step iteration	2946.762	3217.381	3525.726

Table 3
The empirical coverage probabilities of various 90% C.I.s under the 'unstructured' correlation. Number of clusters $n = 200$. Cluster size $m = 3$. Computed based on 500 simulation runs.

Parameters	Gaussian		Poisson		Logistic	
	β_0	β_1	β_0	β_1	β_0	β_1
GEE robust	.880 (.270)	.902 (.210)	.876 (.328)	.888 (.290)	.878 (.359)	.870 (.468)
Bootstrap, percentile, one step iteration	.910 (.298)	.912 (.228)	.896 (.326)	.888 (.288)	.900 (.358)	.888 (.506)
Bootstrap, percentile, full iterations	.916 (.267)	.906 (.208)	.888 (.329)	.906 (.292)	.902 (.361)	.878 (.476)
Bootstrap, hybrid, one step iteration	.904 (.298)	.898 (.228)	.858 (.326)	.868 (.288)	.896 (.358)	.878 (.506)
Bootstrap, hybrid, full iterations	.912 (.267)	.912 (.208)	.894 (.329)	.908 (.292)	.912 (.361)	.886 (.476)

the lengths of the bootstrap C.I.s are shorter than those obtained from the robust methods. The advantages of bootstrap C.I.s become more obvious when the sample size gets larger. Compared with using the working independence, the C.I.s are shorter when the correlation structure is correctly specified while having similar coverage performance. The results for the fourth setup indicate that the bootstrap methods may still work in situations that some conditions of our theory are violated.

Table 2 reports the accumulated computing time for obtaining the bootstrap C.I.s via one-step and full-step method respectively for 500 simulated datasets. As is shown, the computing time for the one-step iterative bootstrap C.I.s are less than that for the full-step iterative bootstrap C.I.s, and this computational saving becomes more significant as the sample size increases. The advantage is less obvious than in the working independence correlation. The reason is that a significant portion of the computation is spent on computing the inverse of the correlation matrix, which is very cheap under working independence.

We further studied the coverage probabilities and interval lengths of various C.I.s under the unstructured working correlation. As argued in [10], the estimation for the $m(m-1)/2$ unknown parameters in the working correlation matrix is only useful when there are few observation times, i.e. m is much smaller than n. Therefore, we let $n = 200$, $m = 3$ in the first three setups and did not consider the fourth setup where $m = n$. The results presented in Table 3 show that in most cases the bootstrap C.I.s have coverage probabilities closer to the nominal level with their lengths slightly wider than the robust C.I.s.

Table 4
The Ohio children wheeze status data (90% confidence intervals).

Method	MS	AGE	MS * AGE
GEE robust	(−.238, −.046)	(−.008,.612)	(−.078,.215)
Bootstrap, percentile, one step iteration	(−.239, −.054)	(−.019,.621)	(−.077,.228)
Bootstrap, percentile, full iterations	(−.244, −.056)	(−.021,.613)	(−.079,.232)
Bootstrap, hybrid, one step iteration	(−.230, −.045)	(−.017,.617)	(−.091,.214)
Bootstrap, hybrid, full iterations	(−.228, −.040)	(−.010,.624)	(−.096,.216)

4.2. The Ohio children wheeze status data

We analyzed a longitudinal dataset which is part of a study of the respiratory health effects of indoor and outdoor air pollution in six US cities. One of the main interests of this study is the effect of maternal smoking on children's respiratory illness. The dataset contains complete records on 537 children from Steubenville, Ohio, each of them being examined annually at ages 7 through 10. Whether the child had a respiratory infection in the year prior to each exam was reported by the mother. The mother's smoking status was recorded at the first interview. This dataset was previously analyzed by Zeger et al. [26].

Let Y_{ij} be the indicator of the ith child's respiration infection status ($1 =$ yes, $0 =$ no) at jth measurement for $1 \leq i \leq 537$ and $1 \leq j \leq 4$. The marginal probability of the respiratory infection, $p_{ij} = P(Y_{ij} = 1|\text{MS}, \text{AGE})$, is modeled as

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 \,\text{MS} + \beta_2 \,\text{AGE} + \beta_3 \,(\text{MS} * \text{AGE}), \tag{37}$$

where MS $=$ 1 if the mother smoked at the first year of the study, 0 otherwise; and AGE is the years since the 9th birthday. Since there is no prior knowledge on the correlation structure, we modeled the association among responses as 'unstructured'. Table 4 summarizes five 90% C.I.s for β_1, β_2 and β_3. The bootstrap C.I.s were constructed using $B = 1000$ resamples of clusters. The average number of iteration steps for the fully iterated bootstrap is 3.608. Table 4 shows that all the C.I.s provide similar results, indicating that MS is significant but neither AGE nor MS-AGE interaction is significant in interpreting the probability of respiratory infection.

Acknowledgments

The first author's research was sponsored by NSF (DMS-0906497, CAREER Award DMS-1151692). The third author's research was partly sponsored by NSF (DMS-0907170), NCI (CA57030), and Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

Appendix

A.1. Useful lemmas

The following Lemmas will be used in the proofs of Theorems 3.1 and 3.3. Lemmas A.1–A.3 are exactly Lemma 3 in [5], Lemma A in [6] and Lemma 4.6 in [17], respectively.

Lemma A.1. *Suppose that*

$$Q_n = o_{P_W}(1) \quad \textit{in } P_Z\textit{-Probability},$$
$$R_n = O_{P_W}(1) \quad \textit{in } P_Z\textit{-Probability}.$$

We have that

$$A_n = o_{P_{ZW}}(1) \Longleftrightarrow A_n = o_{P_W}(1) \quad \textit{in } P_Z\textit{-Probability}, \tag{A.1}$$
$$B_n = O_{P_{ZW}}(1) \Longleftrightarrow B_n = O_{P_W}(1) \quad \textit{in } P_Z\textit{-Probability}, \tag{A.2}$$
$$C_n = Q_n \times O_{P_Z}(1) \Longrightarrow C_n = o_{P_W}(1) \quad \textit{in } P_Z\textit{-Probability}, \tag{A.3}$$
$$D_n = R_n \times O_{P_Z}(1) \Longrightarrow D_n = O_{P_W}(1) \quad \textit{in } P_Z\textit{-Probability}, \tag{A.4}$$
$$E_n = Q_n \times R_n \Longrightarrow E_n = o_{P_W}(1) \quad \textit{in } P_Z\textit{-Probability}. \tag{A.5}$$

Lemma A.2. *Let H be a smooth injection from $\mathbb{R}^p$ to $\mathbb{R}^p$ with $H(x_0) = y_0$. Define $\mathcal{B}_\delta(x_0) = \{x \in \mathbb{R}^p : \|x - x_0\| \leq \delta\}$ and $\partial\mathcal{B}_\delta(x_0) = \{x \in \mathbb{R}^p : \|x - x_0\| = \delta\}$. Then $\inf_{x\in\partial\mathcal{B}_\delta(x_0)} \|H(x) - y_0\| \geq \zeta$ implies $B_\zeta(y_0) = \{y \in \mathbb{R}^p : \|y - y_0\| \leq \zeta\} \subset H(\mathcal{B}_\delta(x_0))$.*

Lemma A.3. *Let $\{m\}$ be a sequence of natural numbers, let $\{a_{mj}\}$ be a triangular array of constants and $\bar{a}_m \equiv \frac{1}{m}\sum_{j=1}^m a_{mj}$. Let B_{mj} be a triangular array of row-exchangeable random variables for $j = 1, \ldots, m$ and $m \in \{m\}$ such that*

$$\frac{1}{m}\sum_{j=1}^{m}(a_{mj}-\bar{a}_m)^2 \to \sigma^2 > 0; \qquad \frac{1}{m}\max_{j\le m}(a_{mj}-\bar{a}_m)^2 \to 0, \tag{A.6}$$

$$\frac{1}{m}\sum_{j=1}^{m}(B_{mj}-\bar{B}_m)^2 \to a^2 > 0 \quad \textit{in probability}, \tag{A.7}$$

$$\lim_{K\to\infty}\limsup_{m\to\infty}\||B_{m1}-\bar{B}_m|1\{|B_{m1}-\bar{B}_m|>K\}\|_2 = 0. \tag{A.8}$$

Then

$$\frac{1}{\sqrt{m}}\sum_{j=1}^{m}(a_{mj}B_{mj}-\bar{a}_m\bar{B}_m) \xrightarrow{d} N(0, a^2\sigma^2).$$

A.2. Proof of Theorem 3.1

We prove the existence of $\widehat{\beta}^*$ by applying Lemma A.2. Consider the mapping $T_n : \beta \mapsto H_n^{-1/2}U_n^*(\beta)$. It is easy to show that T_n is an injection function for $\beta \in B_n(r)$ with probability tending to one based on its derivative

$$\dot{T}_n(\beta) = -\sqrt{n}(H_n/n)^{-1/2}(K_n^*(\beta)/n),$$

and the nonsingularity of V_H and Condition R2′. Define $E_n(r)$ as

$$E_n(r) = \left\{\omega : \|T_n(\beta_0)\| \le \inf_{\beta\in\partial B_n(r)}\|T_n(\beta)-T_n(\beta_0)\|\right\},$$

where $\partial B_n(r) = \{\beta : \|H_n^{1/2}(\beta-\beta_0)\| = (m\tau_n)^{1/2}r\}$. According to Lemma A.2, there exists a $\widehat{\beta}^* \in B_n(r)$ such that $U_n^*(\widehat{\alpha}^*, \widehat{\beta}^*) = 0$ on the set $E_n(r) \cap \{\dot{T}_n(\beta)$ is nonsingular$\}$.

In view of the above discussions, to show the existence of $\widehat{\beta}^*$, we only need to prove $P_{ZW}(E_n(r)) > (1-\epsilon)$ for arbitrarily small ϵ by choosing some proper r as $n \to \infty$. By the Taylor expansion, we have

$$T_n(\beta) - T_n(\beta_0) = -H_n^{-1/2}K_n^*(\bar{\beta}_1)H_n^{-1/2}H_n^{1/2}(\beta-\beta_0),$$

where $\bar{\beta}_1$ lies between β and β_0. So, for $\beta \in \partial B_n(r)$, we have

$$\|T_n(\beta) - T_n(\beta_0)\| \ge (m\tau_n)^{1/2}rz_\lambda^{1/2}, \tag{A.9}$$

where $z_\lambda = \lambda_{\min}(H_n^{-1/2}(K_n^*(\bar{\beta}_1))'H_n^{-1}K_n^*(\bar{\beta}_1)H_n^{-1/2})$. Suppose C is a $p\times p$ matrix. For any $p\times 1$ vector λ with $\|\lambda\| = 1$, we have $\lambda'C'C\lambda \ge (\lambda'C\lambda)^2$ by the Cauchy–Schwarz inequality. Setting $C = H_n^{-1/2}K_n^*(\bar{\beta}_1)H_n^{-1/2}$ in the above and considering Condition R2′, we have

$$P_{ZW}(z_\lambda \ge c_0^2) > 1-\epsilon/2 \tag{A.10}$$

for large enough n. We claim that

$$P_{ZW}(\|T_n(\beta_0)\| \le c_0r(m\tau_n)^{1/2}) \ge 1-\frac{\epsilon}{2} \tag{A.11}$$

for sufficiently large n. Therefore, we have

$$\begin{aligned}
P_{ZW}(E_n(r)) &\ge P_{ZW}(\|T_n(\beta_0)\| \le (m\tau_n)^{1/2}rz_\lambda^{1/2}) \\
&\ge P_{ZW}\left(\{\|T_n(\beta_0)\| \le (m\tau_n)^{1/2}rz_\lambda^{1/2}\} \cap \{z_\lambda \ge c_0^2\}\right) \\
&\ge P_{ZW}(\|T_n(\beta_0)\| \le (m\tau_n)^{1/2}rc_0) + P_Z(z_\lambda \ge c_0^2) - 1 \\
&> 1-\epsilon
\end{aligned}$$

based on (A.9)–(A.11). Since ϵ can be chosen arbitrarily, we complete the proof for the existence of $\widehat{\beta}^*$.

To show (A.11), let S_n be $H_n^{-1/2}U_n^*(\alpha_0, \beta_0)$. Recall that $T_n(\beta_0) = H_n^{-1/2}U_n^*(\widehat{\alpha}^*, \beta_0)$. Then we have

$$\begin{aligned}
P_{ZW}(\|T_n(\beta_0)\| \le c_0r(m\tau_n)^{1/2}) &\ge P_{ZW}(\|S_n\| \le c_0r(m\tau_n)^{1/2}/2) \\
&\quad - P_{ZW}(\|T_n(\beta_0)-S_n\| \ge c_0r(m\tau_n)^{1/2}/2) \\
&\ge I - II.
\end{aligned}$$

By the Chebyshev inequality, we have

$$I \ge 1 - \frac{E_{ZW}\|H_n^{-1/2}U_n^*\|^2}{c_0^2r^2m\tau_n/4} = 1 - \frac{\|W_{n1}\|_2^2 tr(H_n^{-1/2}M_nH_n^{-1/2})}{c_0^2r^2m\tau_n/4} = 1 - \frac{\|W_{n1}\|_2^2 tr(H_n^{-1}M_n)}{c_0^2r^2m\tau_n/4}.$$

Considering the assumed independence between bootstrap weights and original data, we derive the first equality in the above as follows:

$$\begin{aligned} E_{ZW}\|H_n^{-1/2}U_n^*\|^2 &= E_{ZW}\left\|\sum_{i=1}^n W_{ni}H_n^{-1/2}U_{ni}(\alpha_0,\beta_0)\right\|^2 \\ &= E_W W_{n1}^2 E_Z\left\|\sum_{i=1}^n H_n^{-1/2}U_{ni}(\alpha_0,\beta_0)\right\|^2 \\ &= E_W W_{n1}^2 tr\left(\mathrm{Var}_Z\left(\sum_{i=1}^n H_n^{-1/2}U_{ni}(\alpha_0,\beta_0)\right)\right) \\ &= \|W_{n1}\|_2^2 tr(H_n^{-1/2}M_nH_n^{-1/2}). \end{aligned}$$

By noting (18) and choosing $r = (8C/c_0)\sqrt{p/\epsilon}$, we obtain

$$I \geq 1-\epsilon\left(\frac{tr(H_n^{-1}M_n)}{4pm\tau_n}\right) \geq 1-\epsilon\left(\frac{\widetilde{\tau}_n}{4m\tau_n}\right) \geq 1-\frac{\epsilon}{4},$$

where $\widetilde{\tau}_n = \max_{1\leq i\leq n}(\lambda_{\max}(R_i^{-1}(\alpha_0)R_{i0}))$. In the above, the second inequality follows from the expressions of H_n and M_n and the fact that the trace of a square matrix equals to the sum of all its eigenvalues, and the third inequality follows from $\lambda_{\max}\{R_i^{-1}(\alpha_0)R_{i0}\} \leq m_i\lambda_{\max}\{R_i^{-1}(\alpha_0)\}$ which is implied by the fact that $\lambda_{\max}(AB) \leq \lambda_{\max}(A)tr(B)$ for any two positive definite matrices A and B. As for II, we first rewrite $(T_n(\beta_0) - S_n)$ as

$$(H_n/n)^{-1/2}\left(\frac{\partial}{\partial\alpha}|_{\alpha=\bar{\alpha}_1}\frac{U_n^*(\alpha,\beta_0)}{n}\right)\sqrt{n}(\widehat{\alpha}^*-\alpha_0) = A_n\times B_n\times C_n, \tag{A.12}$$

where $\bar{\alpha}_1$ is between $\widehat{\alpha}^*$ and α_0. Note that $A_n\times C_n = O_{P_{ZW}}(1)$ by the root-n consistency of $\widehat{\alpha}^*$ and Lemma A.1. Recall that $U_n^*(\alpha,\beta)$ can be written as $\sum_{i=1}^n W_{ni}D_i'(\beta)V_i^{-1}(\alpha,\beta)S_i(\beta)$. Noting that α is an Euclidean parameter, it follows from Lemma 2.6.15 in [21] that the function class $\{(\partial/\partial\alpha U_{ni}(\alpha,\beta_0) : \alpha\in\mathcal{N}(\alpha_0))\}$ has finite VC index, thereby a Glivenko Cantelli class. Then, by applying the multiplier uniform law of large number, i.e. Lemma 3.6.16 in [21], to $\{(\partial/\partial\alpha)U_{ni}(\alpha,\beta_0) : \alpha\in\mathcal{N}(\alpha_0)\}$, we know that $B_n = o_{P_{ZW}}(1)$. This implies that $(T_n(\beta_0)-S_n) = o_{P_{ZW}}(1)$ based on Lemma A.1. Note that $c_0r(m\tau_n)^{1/2}/2$ is bounded away from zero. Thus, $II \leq \epsilon/4$ for sufficiently large n. Combining the above analysis, we have proved (A.11).

We next focus on the weak consistence of $\widehat{\beta}^*$. The above analysis implies that there exists a $\widehat{\beta}^*$ satisfying $\|H_n^{1/2}(\widehat{\beta}^*-\beta_0)\| \leq (m\tau_n)^{1/2}r$ with probability tending to one. Since $\lambda_{\min}(H_n)/\tau_n\to\infty$ i.e., Condition R1, we have $\tau_n/\lambda_{\min}(H_n)\leq\delta$ for any δ on the event $E_n(r)$. This leads to

$$P_{ZW}(\|\widehat{\beta}^*-\beta_0\| \leq (m\delta)^{1/2}r) \geq P_{ZW}(E_n(r)) > 1-\epsilon.$$

This implies the consistency of $\widehat{\beta}^*$.

To obtain the asymptotic linear expansion of (27), it suffices to show that

$$M_n^{-1/2}H_n(\widehat{\beta}^*-\beta_0) = M_n^{-1/2}\sum_{i=1}^n W_{ni}U_{ni}(\alpha_0,\beta_0) + o_{P_W}(1) \quad \text{in } P_Z\text{-probability}, \tag{A.13}$$

$$M_n^{-1/2}H_n(\widehat{\beta}-\beta_0) = M_n^{-1/2}\sum_{i=1}^n U_{ni}(\alpha_0,\beta_0) + o_{P_Z}(1), \tag{A.14}$$

and then take their difference. We only need to show (A.13) since (A.14) can be viewed as a special case of (A.13), i.e., $W_{ni}=1$ a.s., and $o_{P_Z}(1)$ is also of the order $o_{P_{ZW}}(1)$. To show (A.13), we first apply the intermediate value theorem to the RHS of the first equation, and then to the LHS of the second equation to obtain

$$\begin{aligned} &T_n(\beta_0)-T_n(\widehat{\beta}^*) = H_n^{-1/2}K_n^*(\bar{\beta}_2)(\widehat{\beta}^*-\beta_0) \\ &H_n^{-1/2}U_n^*(\beta_0) = [H_n^{-1/2}K_n^*(\bar{\beta}_2)H_n^{-1/2}]H_n^{1/2}(\widehat{\beta}^*-\beta_0), \\ &H_n^{-1/2}U_n^* = [H_n^{-1/2}K_n^*(\bar{\beta}_2)H_n^{-1/2}]H_n^{1/2}(\widehat{\beta}^*-\beta_0) - \left(\frac{H_n}{n}\right)^{-1/2}\left(\frac{(\partial/\partial\alpha)|_{\alpha=\bar{\alpha}_2}U_n^*(\alpha,\beta_0)}{n}\right)\sqrt{n}(\widehat{\alpha}^*-\alpha_0), \end{aligned}$$

where $\bar{\beta}_2\in B_n(r)$ and $\bar{\alpha}_2$ lies between $\widehat{\alpha}^*$ and α_0. Considering the analysis on (A.12), we have

$$H_n^{-1/2}U_n^* = [H_n^{-1/2}K_n^*(\bar{\beta}_2)H_n^{-1/2}]H_n^{1/2}(\widehat{\beta}^*-\beta_0) + o_{P_{ZW}}(1). \tag{A.15}$$

By Condition R3′, we can further simplify (A.15) to

$$H_n^{1/2}(\widehat{\beta}^*-\beta_0) = H_n^{-1/2}U_n^* + o_{P_{ZW}}(1). \tag{A.16}$$

By multiplying $M_n^{-1/2}H_n^{1/2}$ on both sides of (A.16), we have proved (A.13).

To complete the proof of (28) using the asymptotic linear expansion of (27), we apply Lemma A.3 together with the Cramér–Wold device. To study the conditional weak convergence of

$$M_n^{-1/2}(U_n^* - U_n) = (M_n/n)^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(W_{ni}U_{ni} - U_{ni}),$$

we set $m = n$, $B_{mi} = W_{ni}$ and $a_{mi} = U_{ni}'l$ in Lemma A.3. Obviously, Conditions (A.7) and (A.8) are implied by the bootstrap weight conditions W1–W5; see Lemma 4.7 in [17]. To verify (A.6), we need to show

$$\frac{1}{n}\sum_{i=1}^{n}(U_{ni}'l - U_n'l/n)^2 = \frac{\sum_{i=1}^{n}(U_{ni}'l)^2}{n} - \left(\frac{\sum_{i=1}^{n}U_{ni}'l}{n}\right)^2 \longrightarrow l'V_M l, \tag{A.17}$$

$$\frac{1}{n}\max_{1\le i\le n}(U_{ni}'l - U_n'l/n)^2 \longrightarrow 0. \tag{A.18}$$

The above two convergence results are essentially implied by Conation R4; see Theorem 4 of [25]. Thus, we have proved (28). Moreover, Theorem 2.1 together with Lemma 2.11 in [20] implies that

$$\sup_{z\in\mathbb{R}^p}\left|P_Z(\sqrt{n}(\widehat{\beta} - \beta_0) \le z) - P(N(0, V_G) \le z)\right| = o(1). \tag{A.19}$$

Combining (A.19) with (28), we obtain (29). □

A.3. Proof of Theorem 3.3

Using (36) and the fact that $U_n^*(\widehat{\beta}^*) = 0$, we obtain

$$H_n^{-1/2}H_n^*(\widehat{\beta}_0^*)(\widehat{\beta}_1^* - \widehat{\beta}^*) = [H_n^{-1/2}H_n^*(\widehat{\beta}_0^*)(\widehat{\beta}_0^* - \widehat{\beta}^*)] + H_n^{-1/2}[U_n^*(\widehat{\beta}_0^*) - U_n^*(\widehat{\beta}^*)].$$

We further apply the intermediate value theorem to $U_n^*(\beta)$ to have

$$\begin{aligned} H_n^{-1/2}H_n^*(\widehat{\beta}_0^*)(\widehat{\beta}_1^* - \widehat{\beta}^*) &= \{H_n^{-1/2}[H_n^*(\widehat{\beta}_0^*) - K_n^*(\bar{\beta}^*)]H_n^{-1/2}\}[H_n^{1/2}(\widehat{\beta}_0^* - \widehat{\beta}^*)] \\ &= A_n \times B_n, \end{aligned}$$

where $\bar{\beta}^*$ lies between $\widehat{\beta}_0^*$ and $\widehat{\beta}^*$. Obviously $B_n = O_{P_{ZW}}(1)$ by the $\sqrt{n}$ consistency of $\widehat{\beta}_0^*$ and $\widehat{\beta}^*$ and the limit that $\lim_n H_n/n = V_H$. Lemma 3.6.16 in [21] together with Condition R6 gives

$$\|H_n^*(\widehat{\beta}_0^*)/n - V_H\| \xrightarrow{P_{ZW}} 0 \tag{A.20}$$

based on the consistency of $\widehat{\alpha}^*$ and $\widehat{\beta}_0^*$. This further implies

$$\|(H_n/n)^{-1/2}(H_n^*(\widehat{\beta}_0^*)/n)(H_n/n)^{-1/2} - I\| \xrightarrow{P_{ZW}} 0. \tag{A.21}$$

Note that $\bar{\beta}^* \in B_n(r)$ with probability tending to one. Thus, Condition R3′ and (A.21) imply that $A_n = o_{P_{ZW}}(1)$. So far, we have shown that

$$H_n^{-1/2}H_n^*(\widehat{\beta}_0^*)(\widehat{\beta}_1^* - \widehat{\beta}^*) = o_{P_{ZW}}(1).$$

Further, by considering the limit $(H_n/n) \xrightarrow{P_Z} V_H$, (A.20) and the invertibility of V_H, we complete the whole proof. □

References

[1] R.M. Balan, I. Schiopu-Kratina, Asymptotic results with generalized estimating equations for longitudinal data, Annals of Statistics 33 (2005) 522–541.
[2] P. Barbe, P. Bertail, The Weighted Bootstrap, in: Lecture Notes in Statistics, vol. 98, Springer-Verlag, New York, 1995.
[3] A.C. Cameron, J.B. Gelbach, D.L. Miller, Bootstrap-based improvements for inferences with clustered errors, The Review of Economics and Statistics 90 (2008) 414–427.
[4] S. Chatterjee, A. Bose, Generalized bootstrap for estimating equations, Annals of Statistics 33 (2005) 414–436.
[5] G. Cheng, Z.J. Huang, Bootstrap consistency for general semiparametric M-estimation, Annals of Statistics 38 (2010) 2884–2915.
[6] K. Chen, I. Hu, Z. Ying, Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs, Annals of Statistics 27 (1999) 1155–1163.
[7] C.A. Field, A.H. Welsh, Bootstrapping clustered data, Journal of the Royal Statistical Society: Series B 69 (2007) 369–390.
[8] P. Jassen, J. Jureckova, N. Veraverbeke, Rate of convergence of one- and two-step M-estimators with applications to maximum likelihood and Pitman estimators, Annals of Statistics 25 (1985) 1471–1509.
[9] G. Kauermann, R.J. Carroll, A note on the efficiency of sandwich covariance matrix estimation, Journal of the American Statistical Association 96 (2001) 1387–1396.
[10] K. Liang, S. Zeger, Longitudinal data analysis using generalized linear models, Biometrika 73 (1986) 13–22.

[11] Y. Li, Y. Wang, Smooth bootstrap methods for analysis of longitudinal data, Statistics in Medicine 27 (2007) 937–953.
[12] L.A. Mancl, T.A. DeRouen, A covariance estimator for GEE with improved small-sample properties, Biometrics 57 (2001) 126–134.
[13] D. Mason, M. Newton, A rank statistic approach to the consistency of a general bootstrap, Annals of Statistics 20 (1992) 1611–1624.
[14] L.H. Moulton, S.L. Zeger, Analyzing repeated measures on generalized linear models via the bootstrap, Biometrics 45 (1989) 381–394.
[15] M.C. Paik, Repeated measurement analysis for nonnormal data in small samples, Communications in Statistics, Simulation and Computation 17 (1988) 1155–1171.
[16] D. Pollard, 1990, Empirical Processes: Theory and Applications, NSF-CBMS Regional Conferences Series in Probability and Statistics.
[17] J. Praestgaard, J. Wellner, Exchangeably weighted bootstraps of the general empirical process, Annals of Probability 21 (1993) 2053–2086.
[18] D. Rubin, The Bayesian bootstrap, Annals of Statistics 9 (1981) 130–134.
[19] M. Sherman, S. le Cessie, A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models, Communications in Statistics, Simulation and Communication 26 (1997) 901–925.
[20] A.W. van der Vaart, Asymptotic Statistics, Cambridge University Press, Cambridge, 1998.
[21] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes: With Applications to Statistics, Springer, New York, 1996.
[22] Y.-G. Wang, V. Carey, Working correlation structure misspecification, estimation and covariate design: implications for generalized estimating equations performance, Biometrika 90 (2003) 29–41.
[23] Y.-G. Wang, V. Carey, Unbiased estimating equations from working correlation models for irregularly times repeated measures, Journal of American Statistical Association 99 (2004) 845–853.
[24] R.W.M. Wedderburn, Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method, Biometrika 61 (1974) 439–447.
[25] M. Xie, Y. Yang, Asymptotics for generalized estimating equations with large cluster sizes, Annals of Statistics 31 (2003) 310–347.
[26] S.L. Zeger, K.Y. Liang, P.S. Albert, Models for longitudinal data: a generalized estimating equation approach, Biometrics 44 (1988) 1049–1060.