Review Article

Generalized Estimating Equations in Longitudinal Data Analysis: A Review and Recent Developments

Ming Wang

Division of Biostatistics and Bioinformatics, Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA 17033, USA

Correspondence should be addressed to Ming Wang; mwang@phs.psu.edu

Received 17 March 2014; Revised 29 October 2014; Accepted 16 November 2014; Published 1 December 2014

Academic Editor: Chin-Shang Li

Copyright © 2014 Ming Wang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Generalized Estimating Equation (GEE) is a marginal model popularly applied for longitudinal/clustered data analysis in clinical trials or biomedical studies [1–3]. One longitudinal data example can be taken from a study of orthodontic measurements on children including 11 girls and 16 boys. The response is the measurement of the distance (in millimeters) from the center of the pituitary to the pterygomaxillary fissure, which is repeatedly measured at ages 8, 10, 12, and 14 years. The primary goal is to investigate whether there exists significant gender difference in dental growth measures and the temporal trend as age increases [4]. For such data analysis, it is obvious that the responses from the same individual tend to be "more alike"; thus incorporating within-subject and between-subject variations into model fitting is necessary to improve efficiency of the estimation and the power [5].

There are several simple methods existing for repeated data analysis, that is, ANOVA/MANOVA for repeated measures, but the limitation is the incapability of incorporating covariates. There are two types of approaches, mixed-effect models and GEE [6, 7], which are traditional and are widely used in practice now. Of note is that these two methods have different tendencies in model fitting depending on the study objectives. In particular, the mixed-effect model is an individual-level approach by adopting random effects to capture the correlation between the observations of the same subject [7]. On the other hand, GEE is a population-level approach based on a quasilikelihood function and provides the population-averaged estimates of the parameters [8]. In this paper, we focus on the latter to provide a review and recent developments of GEE. As is well known, GEE has several defining features [9–11]. (1) The variance-covariance matrix of responses is treated as nuisance parameters in GEE and thus this model fitting turns out to be easier than mixed-effect models [12]. In particular, if the overall treatment effect is of primary interest, GEE is preferred. (2) Under mild regularity conditions, the parameter estimates are consistent and asymptotically normally distributed even when the “working” correlation structure of responses is misspecified, and the variance-covariance matrix can be estimated by robust “sandwich” variance estimator. (3) GEE relaxes the distribution assumption and only requires the correct specification of marginal mean and variance as well as the link function which connects the covariates of interest and marginal means.

However, several aspects of GEE are still in controversy since Liang and Zeger [6]. Crowder addressed some issues on inconsistent estimation of within-subject correlation
coefficient under a misspecified “working” correlation structure based on asymptotic theory [7]. In addition, the estimation of the correlation coefficients using the moment-based approach is not efficient; thus the correlation matrix may not be a positive definite matrix in certain cases. Also, Liang and Zeger did not incorporate the constraints on the range of correlation which was restricted by the marginal means because the estimation of the correlation coefficients was simply based on Pearson residuals [6]. Chaganty and Joe discussed this issue for dependent Bernoulli random variables [13], and later Sabo and Chaganty made future explanation [14]. For example, Sutradhar and Das pointed out under misspecification the correlation coefficient estimates did not converge to the true values [15]. Furthermore, for discrete random vectors, the correlation matrix was usually complicated, and it was not easy to attain multivariate distributions with specified correlation structures. These limitations lead researchers to actively work on this area to develop novel methodologies. Several alternative approaches for estimating the correlation coefficients have been proposed; for example, one method was based on “Gaussian” estimation [16, 17], and the basic idea was to estimate the correlation coefficients based on multivariate normal estimating equations, and the feature was that this estimation can ensure the estimated correlation matrix was positive-definite. Wang and Carey proposed to estimate the correlation coefficients by differentiating the Cholesky decomposition of the working correlation matrix [18]. Also, Qu and Lindsay (2003) proposed similar Gaussian or quadratic estimating equations [19]. In particular, for binary longitudinal data, the estimation of the correlation coefficients was proposed based on conditional residuals [20–22]. Nevertheless, in this paper, the above issues are not discussed in great depth, and the assumption that, under the regular mild conditions, the consistency of parameter estimates as well as within-subject correlation coefficient estimate holds is satisfied. Thus, three specific topics including model selection, power analysis, and the issue of informative cluster size are mainly focused on and the recent developments are reviewed in the following sections.

2. Method

2.1. Notation and GEE. Suppose that longitudinal/clustered data consists of $K$ subjects/clusters. For subject/cluster $i$ ($i = 1, 2, \ldots, K$), suppose that there are $n_i$ observations and $Y_{ij}$ denotes the $j$th response ($j = 1, \ldots, n_i$), and let $X_{ij}$ denote a $p \times 1$ vector of covariates. Let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ denote the response vector for the $i$th subject with the mean vector noted by $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i})'$ where $\mu_{ij}$ is the corresponding $j$th mean. The responses are assumed to be independent across subjects/clusters but correlated within each subject/cluster. The marginal model specifies that a relationship between $\mu_{ij}$ and the covariates $X_{ij}$ is written as follows:

$$g(\mu_{ij}) = X_{ij}'\beta,$$

where $g$ is a known link function and $\beta$ is an unknown $p \times 1$ vector of regression coefficients with the true value $\beta_0$. The conditional variance of $Y_{ij}$ given $X_{ij}$ is specified as $\text{Var}(Y_{ij} | X_{ij}) = \nu(\mu_{ij})\phi$, where $\nu$ is a known variance function of $\mu_{ij}$ and $\phi$ is a scale parameter which may need to be estimated. Mostly, $\nu$ and $\phi$ depend on the distributions of outcomes. For instance, if $Y_{ij}$ is continuous, $\nu(\mu_{ij})$ is specified as 1, and $\phi$ represents the error variance; if $Y_{ij}$ is count, $\nu(\mu_{ij}) = \mu_{ij}$, and $\phi$ is equal to 1. Also, the variance-covariance matrix for $Y_i$ is noted by $V_i = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}$, where $A_i = \text{Diag}(\nu(\mu_{i1}), \ldots, \nu(\mu_{in_i}))$ and the so-called “working” correlation structure $R_i(\alpha)$ describes the pattern of measures within subject, which is of size $n_i \times n_i$ and depends on a vector of association parameters denoted by $\alpha$. Table 1 provides summary of commonly used “working” correlation structures with the moment-based estimates for $\alpha$ (more details in http://www.okstate.edu/sas/). Note that the iterative algorithm is applied for estimating $\alpha$ using the Pearson residuals $e_{ij} = (Y_{ij} - \mu_{ij})/\sqrt{\nu(\mu_{ij})}$ calculated from the current value of $\beta$. Also, the scale parameter $\phi$ can be estimated by

$$\hat{\phi} = \frac{1}{N - p} \sum_{i=1}^{K} \sum_{j=1}^{n_i} e_{ij}^2,$$

where $N = \sum_{i=1}^{K} n_i$ is the total number of observations and $p$ is covariates dimensionality.

Based on Liang and Zeger [6], GEE yields asymptotically consistent $\hat{\beta}$ even when the “working” correlation structure ($R_i(\alpha)$) is misspecified, and the estimate of $\beta$ is obtained by solving the following estimating equation:

$$U(\beta) = \sum_{i=1}^{K} D_i' V_i^{-1} (Y_i - \mu_i) = 0,$$

where $D_i = \partial\mu_i / \partial\beta'$. Under mildregularity conditions, $\hat{\beta}$ is asymptotically normally distributed with a mean $\beta_0$ and a covariance matrix estimated based on the sandwich estimator

$$\hat{V}_{LZ} = \left( \sum_{i=1}^{K} D_i' V_i^{-1} D_i \right)^{-1} \hat{M}_{LZ} \left( \sum_{i=1}^{K} D_i' V_i^{-1} D_i \right)^{-1},$$

with

$$\hat{M}_{LZ} = \sum_{i=1}^{K} D_i' V_i^{-1} \text{Cov}(Y_i) V_i^{-1} D_i,$$

by replacing $\alpha$, $\beta$, and $\phi$ with their consistent estimates, where $\text{Cov}(Y_i) = \hat{\rho}(i)'$, with $\hat{\rho}(i) = Y_i - \hat{\mu}_i$ is an estimator of the variance-covariance matrix of $Y_i$ [6, 23]. This “sandwich” estimator is robust in that it is consistent even if the correlation structure ($V_i$) is misspecified. Note that if $V_i$ is correctly specified, then $\hat{V}_{LZ}$ reduces to ($\sum_{i=1}^{K} D_i' V_i^{-1} D_i)^{-1}$, which is often referred to as the model-based variance estimator [24]. Thus, a Wald Z-test can be performed based on asymptotic normal distribution of the test statistic. Next, we will overview model selection criteria and particularly “working” correlation structure selection criteria with regard to GEE.
### Table 1: Summary of commonly used “working” correlation structures for GEE.

<table>
<thead>
<tr>
<th>Correlation structure</th>
<th>Corr($Y_{ij}, Y_{ik}$)</th>
<th>Sample matrix</th>
<th>Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent</td>
<td>$\begin{cases} 1 &amp; j = k \ 0 &amp; j \neq k \end{cases}$</td>
<td>$\begin{pmatrix} 1 &amp; 0 \ 0 &amp; 0 \end{pmatrix}$</td>
<td>NA</td>
</tr>
<tr>
<td>Exchangeable</td>
<td>$\begin{cases} 1 &amp; j = k \ \alpha &amp; j \neq k \end{cases}$</td>
<td>$\begin{pmatrix} 1 &amp; \alpha \ \alpha &amp; 1 \end{pmatrix}$</td>
<td>$\alpha = \frac{1}{(N' - p)} \sum_{i=1}^{K} \sum_{j \neq k} e_{ij} f_{ik}$</td>
</tr>
<tr>
<td>$k$-dependent</td>
<td>$\begin{cases} 1 &amp; m = 0 \ \alpha_m &amp; m = 1, 2, \ldots, k \ 0 &amp; m &gt; k \end{cases}$</td>
<td>$\begin{pmatrix} 1 &amp; \alpha_1 &amp; 0 \ \alpha_1 &amp; 1 &amp; \alpha_1 \ 0 &amp; \alpha_1 &amp; 1 \end{pmatrix}$</td>
<td>$\alpha_m = \frac{1}{(K_m - p)} \sum_{i=1}^{K} \sum_{j \neq k} e_{ij} f_{ij}$</td>
</tr>
<tr>
<td>Autoregressive AR(1)</td>
<td>$\begin{cases} 1 &amp; m = 0 \ \alpha_m &amp; m = 1, 2, \ldots, n_i - j \end{cases}$</td>
<td>$\begin{pmatrix} 1 &amp; \alpha_1 &amp; 0 \ \alpha_1 &amp; 1 &amp; \alpha_1 \ \alpha_2 &amp; \alpha_1 &amp; 1 \end{pmatrix}$</td>
<td>$\alpha = \frac{1}{(N' - p)} \sum_{i=1}^{K} \sum_{j \neq k} e_{ij} f_{ik}$</td>
</tr>
<tr>
<td>Toeplitz</td>
<td>$\begin{cases} 1 &amp; m = 0 \ \alpha_m &amp; m = 1, 2, \ldots, n_i - j \end{cases}$</td>
<td>$\begin{pmatrix} 1 &amp; \alpha_1 &amp; \alpha_2 \ \alpha_2 &amp; 1 &amp; \alpha_1 \ \alpha_3 &amp; \alpha_2 &amp; 1 \end{pmatrix}$</td>
<td>$\alpha = \frac{1}{(K - p)} \sum_{i=1}^{K} \sum_{j \neq k} e_{ij} f_{ik}$</td>
</tr>
<tr>
<td>Unstructured</td>
<td>$\begin{cases} 1 &amp; j = k \ \alpha_{jk} &amp; j \neq k \end{cases}$</td>
<td>$\begin{pmatrix} 1 &amp; \alpha_{12} &amp; \alpha_{13} \ \alpha_{21} &amp; 1 &amp; \alpha_{23} \ \alpha_{31} &amp; \alpha_{32} &amp; 1 \end{pmatrix}$</td>
<td>$\alpha = \frac{1}{(K - p)} \sum_{i=1}^{K} e_{ij} f_{ik}$</td>
</tr>
</tbody>
</table>

2.2. Model Selection of GEE. In this section, we will discuss the model selection criteria available of GEE. There are several reasons why model selection of GEE models is important and necessary: (1) GEE has gained increasing attention in biomedical studies which may include a large group of predictors [25–28]. Therefore, variable selection is necessary for determining which are included in the final regression model by identifying significant predictors; (2) it is already known that one feature of GEE is that the consistency of parameter estimates can still hold even when the “working” correlation structure is misspecified. But, correctly specifying “working” correlation structure can definitely enhance the efficiency of the parameter estimates in particular when the sample size is not large enough [16, 24, 25, 29]. Therefore, how to select intrasubject correlation matrix plays a vital role in GEE with improved finite-sample performance; (3) the variance function $\nu(\mu)$ is another potential factor affecting the goodness-of-fit of GEE [25, 30]. Correctly specified variance function can assist in the selection of covariates and an appropriate correlation structure [31, 32]. Different criteria might be needed due to the goal of model selection [24, 29, 33], and next I will particularly introduce the existing approaches on the selection of “working” correlation structure with its own merits and limitations [34].

According to Rotnitzky and Jewell, the adequacy of “working” correlation structure can be examined through the selection of “working” correlation structure with its own merits and limitations [34].

$$\Gamma = (\sum_{i=1}^{K} D_i' V_i^{-1} D_i)^{-1} \tilde{M}_{LZ},$$

where $\tilde{M}_{LZ}$ has been defined in Section 2.1 [35]. The statistic $RJ(R)$ is defined by

$$RJ(R) = \sqrt{(1 - RJ1)^2 + (1 - RJ2)^2},$$

where $RJ1 = \text{trace}(\Gamma) / p$ and $RJ2 = \text{trace}(\Gamma^2) / p$, respectively.

If the “working” correlation structure $R$ is correctly specified, $RJ1$ and $RJ2$ will be thus close to 1, leading to $RJ(R)$ approaching 0. Thus, $RJ1$, $RJ2$, and $RJ(R)$ can all be used for correlation structure selection.

Shults and Chaganty [36] proposed a criterion for selecting “working” correlation structure based on the minimization of the generalized error sum of squares (ESS) given as follows:

$$\text{ESS}(\alpha, \beta) = \sum_{i=1}^{K} (Y_i - u_i)' V_i^{-1} (Y_i - u_i)$$

$$= \sum_{i=1}^{K} Z_i' (\beta) R_i^{-1} (\alpha) Z_i (\beta),$$

where $Z_i(\beta) = A^{1/2}(Y_i - u_i)$. The criterion is defined by

$$SC = \frac{\text{ESS}(\alpha, \beta)}{(N - p - q)},$$

where $SC$ is the sample correlation coefficient.
where \( N = \sum_{i=1}^{K} n_i \) is the total number of observations, \( p \) is the number of regression parameters, and \( q \) is the number of correlation coefficients within the “working” correlation structure. Another extended criterion from SC was proposed by Carey and Wang [37], where the Gaussian pseudolikelihood (GP) is adopted, and it is given by

\[
GP(R) = -0.5 \times \sum_{i=1}^{K} (Y_i - u_i)' V_i^{-1} (Y_i - u_i) + \log (|V_i|), \quad (9)
\]

where a better “working” correlation structure yields a larger GP. In their work, they also showed that GP criterion held better performance than R) via simulation.

Another criterion is proposed by Pan [38], which modified Akaike information criterion (AIC) [39] in adaption to GEE. Due to the fact that GEE is not likelihood-based, thus it is called quasi-likelihood under the independence model criterion (QIC) [40]. The basic idea is to calculate the expected Kullback-Leibler discrepancy using the quasi-likelihood under the independence “working” correlation structure simultaneously [41, 42], but it is called quasi-likelihood under the independence “working” correlation structures because, in certain cases, the same assumption due to the lack of a general and tractable quasi-likelihood for the correlated data under any other complex “working” correlation structures. QIC(R) is defined by

\[
QIC(R) = -2 \Psi (\hat{\beta}(R); I) + 2 \text{trace} \left( \hat{\Omega}_L \tilde{V}_{LZ} \right), \quad (10)
\]

where the quasilikelihood \( \Psi(\hat{\beta}(R); I) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} Q(\beta(R), \hat{\phi}; \{Y_{ij}, X_{ij}\}) \) with \( Q(\mu, \phi; y) = \int_y (y - t) \phi(D(t)) dt \) defined by [12], \( \hat{\beta} \) and \( \hat{\phi} \) are obtained under the hypothesized “working” correlation structure \( R, \hat{\Omega}_L = \sum_{i=1}^{K} D_i' V_i^{-1} D_i |_{\beta=\hat{\beta}, \phi=\hat{\phi}}, \) and \( \tilde{V}_{LZ} \) is defined above with replacement of \( \beta \) by \( \hat{\beta}(R) \) [38]. Note that, in this work, Pan ignored the second term in Taylor’s expansion of the discrepancy and showed its influence was not substantial among his simulation set-ups. Later on, Hardin and Hilbe (2003) made slight modification on QIC(R) by using \( \hat{\beta}(I) \) and \( \hat{\phi}(I) \) for more stability, and QIC(R)_{HIC} is given by

\[
QIC(R)_{HIC} = -2 \Psi (\hat{\beta}(I); I) + 2 \text{trace} \left( \hat{\Omega}_L \tilde{V}_{LZ} \right). \quad (11)
\]

Note that QIC(R) and QIC(R)_{HIC} do not perform well in distinguishing the independence and exchangeable “working” correlation structures because, in certain cases, the same regression parameter estimates can be obtained under these two structures. Also, the attractive property of the QIC criterion is that it allows the selection of the covariates and “working” correlation structure simultaneously [41, 42], but this measure is more sensitive to the mean structure because QIC is particularly impacted by the first term and the second term which plays a role as a penalty. To better select “working” correlation structure, Hin and Wang proposed correlation information criterion (CIC) defined by

\[
CIC = \text{trace} \left( \hat{\Omega}_L \tilde{V}_{LZ} \right). \quad (12)
\]

In their work, CIC was shown to outperform QIC when the outcomes were binary through simulation studies [43]. One limitation of this criterion is that it cannot penalize the overparameterization; thus the performance is not well in comparison with two correlation structures having quite different numbers of correlation parameters.

Another attractive criterion is the extended quasilikelihood information criterion (EQIC) proposed by Wang and Hin [25] by using the extended quasilikelihood (EQL) defined by Nelder and Pregibon based on the deviance function, which is shown below under the independent correlation structure [44]:

\[
Q^*(\beta, \phi; I) = -\frac{1}{2\phi} D(\beta,I) - \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \log \left( 2\pi \phi A(\mu_{ij}) \right), \quad (13)
\]

where the sum of deviances \( D(\beta,I) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} -2\phi(Q(y_{ij}; \mu_{ij}) - Q(y_{ij}, y_{ij})) \) with Q(\cdot) being the quasilikelihood defined as above. Therefore, EQIC is defined by

\[
EQIC(R) = -\frac{1}{\phi} D(\beta,I) + \sum_{i=1}^{K} \sum_{j=1}^{n_i} \log \left( 2\pi \phi A(\mu_{ij}) \right) + 2 \text{trace} \left( \hat{\Omega}_L \tilde{V}_{LZ} \right), \quad (14)
\]

where some adjustments were applied to \( A(\mu) \) by adding a small constant \( k \) with the optimal chosen value as 1/6. The author indicated that the covariates were first selected based on QIC, and the variance function could be identified as the one minimizing EQIC given the selected covariates; then “working” correlation structure selection could be achieved based on CIC; in addition, they found out that the covariates selection by EQIC given different working variance functions was more consistent than that based on QIC [45].

Besides those criteria mentioned above, Cantoni et al. also discussed the covariate selection for longitudinal data analysis [46]; also, a variance function selection was mentioned by Pan and Mackenize [30] as well as Wang and Lin [47]; in addition, more work on “working” correlation structure selection was addressed by Chaganty and Joe [48], Wang and Lin [47], Gosho et al. [49, 50], Jang [51], Chen [52], and Westgate [53–55], among others. Overall, the model selection of GEE is nontrivial, where the best selection criterion is still being pursued [56], and the recent work by Wang et al. can be followed up as the rule of thumb [45].

2.3. Sample Size and Power of GEE. It is well known that the calculation of sample size and power is necessary and important for planning a clinical trial, which have been well studied for independent observations [1]. With the wide applications of GEE in clinical trials, this topic for correlated/clustered data has gained more attention than ever [5, 57]. The general method for sample size/power calculated was discussed by Liu and Liang [58], where the generalized score test was utilized to draw statistical inference and the resulting noncentral chi-square distribution of test statistic under the alternative hypothesis was derived; however, in some special cases, that is, correlated binary data with nonexchangeable correlation structure, there was no close
form available along the outline of that formula. Afterwards, Shih provided an alternative formula on sample size/power calculation, which relied on Wald tests using the estimates of regression parameters and robust variance estimators [59]. For example, in a study with one parameter of interest $\beta$, the hypothesis of interest can be formulated as

$$H_0: \beta = 0 \quad \text{versus} \quad H_a: \beta = b \neq 0,$$

where $b$ is the expected value. Thus, based on a two-sided $Z$-test with type I error $\eta$, the power denoted by $\delta$ can be obtained by

$$\delta = 1 - \Phi \left( \frac{Z_{\eta/2} - b \sqrt{K}}{\sqrt{\hat{v}_{KL}}} \right),$$

where $K$ is sample size and $\hat{v}_{KL}$ is the robust variance estimator corresponding to $\beta$ in the estimate of $K\hat{V}_{LZ}$. Accordingly, the sample size is given by

$$K = \frac{\hat{v}_{KL} (Z_{\eta/2} - Z_{1-\delta})^2}{b^2}.$$  \hfill (17)

For correlated continuous data, the calculation is straightforward using (16); however, in particular, for correlated binary data, more work will be needed [60], and Pan provided explicit formulas for $\hat{v}_{KL}$ under various situations as follows [61]:

$$\hat{v}_{KL} = \Omega \left[ \frac{1}{n \hat{p}_0 (1-\hat{p}_0)} + \frac{1}{(1-\hat{p}) \hat{p}_1 (1-\hat{p}_1)} \right],$$

where $\Omega = K(\sum_{i=1}^K n_i^{-1} R_i^{-1} V_i R_i^{-1} 1 n_i)/(\sum_{i=1}^K n_i^{-1} R_i^{-1} 1 n_i)^2$ with $n$ as the proportion of subjects assigned to the control group and $\hat{p}_0$ and $\hat{p}_1$ as the mean for control and case groups [61]. The detailed calculations of $\hat{v}_{KL}$ under several important special cases are given by

If $R_i = V_i = CS: \Omega = \frac{K}{\sum_{i=1}^K (n_i / (1 + (n_i - 1) \alpha))}$;
If $R_i = I, V_i = CS: \Omega = \frac{K \sum_{i=1}^K n_i [1 + (n_i - 1) \alpha]}{(\sum_{i=1}^K n_i)^2}$;
If $R_i = V_i = AR (1): \Omega = \frac{K (1 + \alpha)}{\sum_{i=1}^K [n_i - (n_i - 2) \alpha]}$;
If $R_i = I, V_i = AR (1): \Omega = \frac{K \sum_{i=1}^K [n_i + 2 (n_i - 1) \alpha + 2 (n_i - 2) \alpha^2 + \cdots + 2 \alpha^{n_i-1}]}{\sum_{i=1}^K n_i^2}.$

These formulas can be directly used in practice, which has covered most situations encountered in clinical trials [61]. Note that when $R_i = V_i = CS$, Liu and Liang (1997) provided a different formula of sample size compared with (17) with $n_i = n$, which is

$$K = \left( \frac{(Z_{1-\eta/2} + Z_{1-\delta})^2}{\delta^2} \times [(1-\hat{p}) \hat{p}_0 (1-\hat{p}_0) + \pi \hat{p}_1 (1-\hat{p}_1)] \times [1 + (n-1) \alpha] \right)^{-1}.$$  \hfill (20)

Be aware that the difference is due to the test methods, the Wald $Z$-test used by Pan [61] and the score test applied by Liu and Liang [58]. Note that, in some cases, the score test may be preferred [62]. Although some other works exist for sample size/power calculation, they focused on the other alternative approaches rather than GEE [63, 64]; thus we do not discuss them here. For correlated Poisson data, the sample size/power calculation is more challenging due to the occurrence of overdispersion or sparsity, where negative binomial regression model may be explored [62, 65–67].

On the other hand, there are several concerns [68]. First, we here focus on the calculation of the sample size $K$ assuming $n_i$ is known; however, based on the power formula (16), $\hat{v}_{KL}$ depends on $n_i$ and thus increasing $n_i$ can also assist in power improvement but turns out to be less effective than $K$ [69]. Second, the sample size/power calculation may be restricted to the limitation of clusters, for example, clustered randomized trials (CRTs), where the number of clusters could be relatively small. For example, by the literature review of published CRTs, the median number of clusters is shown as 21 [70]. In such situations, the power formula adjusted for the small samples in GEE is necessary, which has drawn attention from researchers recently [71–75].

2.4. Clustered Data with Informative Cluster Size. The application of GEE in clustered data with informative cluster size is another special topic [76]. Taking an example of a periodontal disease study, the number of teeth for each patient may be related to the overall oral health of the individual; in other words, the worse the oral health is, the less the number of teeth is and, thus, cluster size may influence the distribution of the oral outcomes, which is called informative cluster size [45, 77]. Such issues commonly occur in biomedical studies (e.g., genetic disease studies), and rigorous statistical methods are needed for valid statistical inference [78]. Note that if the maximum of cluster size exists and is known, then this can be treated as (informative) missing data problem, which can be solved via the weighted estimating equations proposed by Robins et al. [79]; however, if the maximum is unknown or not accessible, the method of within-cluster resampling (WCR) proposed by Hoffman et al. could be applied [80]. The basic idea is that, for each of $L$ resampled replicate data based on a Monte Carlo method ($L$ is a large number, i.e., 10,000), one observation is randomly extracted from each cluster, where $\hat{\beta}_1$ with variance estimator $\hat{\Sigma}$ can be obtained from a regular score equation denoted by $S_1(\hat{\beta})$ for independent observations (i.e., linear regression for
continuous data; logistic regression for binary data; Poisson regression for count data), $l = 1, 2, \ldots, L$. The details are shown as follows:

$$S_l(\hat{\beta}) = \sum_{i=1}^{K} S_{ij}(\hat{\beta}) I \{ j \in r_l \} = 0;$$

$$\tilde{\beta}_{wcr} = \frac{1}{L} \sum_{l=1}^{L} \hat{\beta}_j;$$

$$\tilde{v}_{wcr} = \frac{1}{L} \sum_{l=1}^{L} S_{ij} - \frac{1}{L} \sum_{l=1}^{L} (\hat{\beta}_j - \tilde{\beta}_{wcr}) (\hat{\beta}_j - \tilde{\beta}_{wcr})^T,$$

where $S_{ij}(\hat{\beta}) = X_{ij}'V_{ij}^{-1}(Y_{ij} - X_{ij}'\hat{\beta})$ with $r_l$ as the set of data index selected from the $i$th cluster in $l$th replicate data. Alternatively, the approach considered by Williamson et al. by adopting the weighted estimating equations performs asymptotically equivalently as WCR and also avoids intensive computing, and it is referred to as the cluster-weighted GEE (CWGEE) [81]. The estimating equation is

$$S(\hat{\beta}) = \sum_{i=1}^{K} \frac{1}{K} S_{ij}(\hat{\beta}) = 0,$$

where $S_{ij}$ is defined as the same as above, but what is different is that the subscription $j$ ranges from 1 to $n_l$, not restricted by the index $r_l$. Note that as $L \to \infty$, $(1/L) \sum_{l=1}^{L} S_{ij}(\hat{\beta})$ converges to its expected estimating function and is asymptotically equivalent to $S(\beta)$.

This method was also explored or extended for the correlated data with nonignorable cluster size by Benhin et al. and Cong et al. [82, 83]. Furthermore, a more efficient method called modified WCR (MWCR) was proposed by Chiang and Lee, where minimum cluster size $n_l > 1$ subjects were randomly sampled from each cluster, and then GEE models for balanced data were applied for estimation by incorporating the intracluster correlation; thus MWCR might be a more efficient way for analysis [84]. But MWCR is not always satisfactory and Pavlou et al. recognized the sufficient conditions of the data structure and the choice of “working” correlation structure, which allowed the consistency of the estimates from MWCR [85]. In addition, Wang et al. extended the above work to the clustered longitudinal data, which are collected as repeated measures on subjects arising in clusters, with potential informative cluster size [45]. Examples include health studies of subjects from multiple hospitals or families. With the adoption and comparison of GEE, WCR, and CWGEE, the author claimed that CWGEE was recommended because of the comparable performance with WCR and the lack of intensive Monte Carlo computation in terms of well preserved coverage rates and desirable power properties, while GEE models led to invalid inference due to the biased parameter estimates via extensive simulation studies and real data application of a periodontal disease study [45]. In addition, for observed-cluster inference, Seaman et al. discussed the methods, including weighted and doubly weighted GEE and the shared random-effects models for comparison, and showed the conditions under which the shared random-effects model described members with observed outcomes $Y$ [86]. More work can be found in [87–90], among others.

3. Simulation

In this section, we focus on “working” correlation structure selection and compare the performances of the existing criteria through simulation studies. Two types of outcomes are considered, continuous and count responses. The models for data generation are as follows:

$$u_{ij} = \beta_0 + \beta_1 \times x_{ij};$$

$$\log(u_{ij}) = \beta_0 + \beta_1 \times x_{ij},$$

where $\beta_0 = \beta_1 = 0.5, i = 1, 2, \ldots, I$ with $I = 50, 100, 200, 500$ and $j = 1, 2, \ldots, J$ with $J = 4, 8$. The covariates $x_{ij}$ are i.i.d. from a standard uniform distribution $\text{Unif}(0,1)$. For each scenario, we generate the data based on the underlying true correlation structures as independent (IND), exchangeable (EXCH), and autoregressive (AR-1) with $\alpha = 0.3, 0.7.$ 1,000 Monte Carlo data sets are generated for each scenario, where the estimates of regression parameters and within-subject correlation matrix and seven model selection criteria measures are calculated using the “working” correlation structure of IND, EXCH, and AR-1. The partial simulation results are provided in Tables 2, 3, and 4, where the results of CIC are not shown because they are the same as those of QIC.

Based on the results, RJ does not perform well for the scenarios with either continuous or binary outcomes, while RJ1 and RJ2 have comparable performances and can select the true underlying correlation structure in most scenarios with better performance under large sample size. QIC is not satisfactory when the true correlation structure is independent but has advantageous performance for the scenarios with the true correlation structure as exchangeable or AR-1. On the other hand, SC and GP do not perform well for longitudinal data with normal responses, but the performance is slightly improved for longitudinal data with binary outcomes. The results may vary due to variety of factors including the types of “working” correlation structure considered for model fitting, the sample size, and/or the magnitude of correlation coefficient. For the future work, there is a necessity to find out a robust criterion for “working” correlation structure selection of GEE, and more advanced approaches are emerging currently.

4. Future Direction and Discussion

In this paper, we provide a review of several specific topics such as model selection with emphasis on the selection of “working” correlation structure, sample size and power calculation, and clustered data analysis with informative cluster size related to GEE for longitudinal/correlated data. The simulation studies are conducted for providing numerical comparisons among five types of model selection criteria [91, 92]. Until now, novel methodologies are still needed and being developed due to the increasing usage and potential
Table 2: Simulation for longitudinal data with independent correlation matrix.

<table>
<thead>
<tr>
<th>n</th>
<th>K</th>
<th>Criterion</th>
<th>IND</th>
<th>EXCH</th>
<th>AR-1</th>
<th>IND</th>
<th>EXCH</th>
<th>AR-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Normal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
<td>QIC</td>
<td>198</td>
<td>393</td>
<td>409</td>
<td>202</td>
<td>374</td>
<td>424</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ</td>
<td>327</td>
<td>423</td>
<td>250</td>
<td>312</td>
<td>421</td>
<td>267</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ1</td>
<td>388</td>
<td>322</td>
<td>290</td>
<td>399</td>
<td>316</td>
<td>285</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ2</td>
<td>384</td>
<td>327</td>
<td>289</td>
<td>388</td>
<td>320</td>
<td>292</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>SC</td>
<td>488</td>
<td>1</td>
<td>512</td>
<td>351</td>
<td>310</td>
<td>339</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>GP</td>
<td>547</td>
<td>0</td>
<td>453</td>
<td>368</td>
<td>306</td>
<td>326</td>
</tr>
</tbody>
</table>

| 8  |     | QIC       | 209 | 377   | 414   | 185 | 407   | 408  |
|    |     | RJ        | 338 | 415   | 247   | 340 | 410   | 250  |
|    |     | RJ1       | 389 | 349   | 262   | 381 | 358   | 261  |
|    |     | RJ2       | 389 | 353   | 258   | 372 | 357   | 271  |
|    | 100 | SC        | 482 | 1     | 517   | 352 | 346   | 302  |
|    |     | GP        | 520 | 0     | 480   | 360 | 348   | 292  |

Table 3: Simulation for longitudinal data with exchangeable correlation matrix with $\alpha = 0.3$.

<table>
<thead>
<tr>
<th>n</th>
<th>K</th>
<th>Criterion</th>
<th>IND</th>
<th>EXCH</th>
<th>AR-1</th>
<th>IND</th>
<th>EXCH</th>
<th>AR-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Normal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
<td>QIC</td>
<td>106</td>
<td>699</td>
<td>195</td>
<td>53</td>
<td>758</td>
<td>189</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ</td>
<td>419</td>
<td>139</td>
<td>442</td>
<td>869</td>
<td>5</td>
<td>126</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ1</td>
<td>0</td>
<td>963</td>
<td>37</td>
<td>12</td>
<td>898</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ2</td>
<td>0</td>
<td>959</td>
<td>41</td>
<td>22</td>
<td>876</td>
<td>102</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>SC</td>
<td>0</td>
<td>593</td>
<td>407</td>
<td>282</td>
<td>650</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>GP</td>
<td>1</td>
<td>593</td>
<td>406</td>
<td>412</td>
<td>524</td>
<td>64</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>QIC</td>
<td>31</td>
<td>879</td>
<td>90</td>
<td>7</td>
<td>867</td>
<td>126</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ</td>
<td>350</td>
<td>88</td>
<td>562</td>
<td>911</td>
<td>2</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ1</td>
<td>0</td>
<td>995</td>
<td>5</td>
<td>2</td>
<td>946</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ2</td>
<td>0</td>
<td>996</td>
<td>4</td>
<td>10</td>
<td>933</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>SC</td>
<td>0</td>
<td>598</td>
<td>402</td>
<td>339</td>
<td>635</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GP</td>
<td>0</td>
<td>501</td>
<td>499</td>
<td>445</td>
<td>531</td>
<td>24</td>
</tr>
</tbody>
</table>

| 8  |     | QIC       | 80  | 828   | 92    | 50  | 876   | 74   |
|    |     | RJ        | 10  | 395   | 595   | 813 | 6     | 181  |
|    |     | RJ1       | 0   | 1000  | 0     | 0   | 987   | 13   |
|    |     | RJ2       | 0   | 1000  | 0     | 0   | 966   | 25   |
|    | 100 | SC        | 0   | 488   | 513   | 302 | 696   | 2    |
|    |     | GP        | 0   | 511   | 489   | 497 | 500   | 3    |
| 8  |     | QIC       | 17  | 953   | 30    | 8   | 973   | 19   |
|    |     | RJ        | 0   | 408   | 592   | 861 | 0     | 139  |
|    |     | RJ1       | 0   | 1000  | 0     | 0   | 997   | 3    |
|    |     | RJ2       | 0   | 1000  | 0     | 0   | 993   | 7    |
|    | 100 | SC        | 0   | 470   | 530   | 328 | 672   | 0    |
|    |     | GP        | 0   | 526   | 474   | 486 | 514   | 0    |
Table 4: Simulation for longitudinal data with AR-1 correlation matrix with $\alpha = 0.3$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$K$</th>
<th>Criterion</th>
<th>IND</th>
<th>EXCH</th>
<th>AR-1</th>
<th>IND</th>
<th>EXCH</th>
<th>AR-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>8</td>
<td>QIC</td>
<td>91</td>
<td>166</td>
<td>743</td>
<td>66</td>
<td>170</td>
<td>764</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ</td>
<td>712</td>
<td>142</td>
<td>146</td>
<td>925</td>
<td>12</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ1</td>
<td>0</td>
<td>478</td>
<td>522</td>
<td>7</td>
<td>505</td>
<td>488</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ2</td>
<td>0</td>
<td>466</td>
<td>534</td>
<td>20</td>
<td>499</td>
<td>481</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SC</td>
<td>0</td>
<td>480</td>
<td>520</td>
<td>220</td>
<td>350</td>
<td>430</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>GP</td>
<td>0</td>
<td>543</td>
<td>457</td>
<td>303</td>
<td>332</td>
<td>365</td>
</tr>
<tr>
<td></td>
<td></td>
<td>QIC</td>
<td>25</td>
<td>116</td>
<td>859</td>
<td>7</td>
<td>122</td>
<td>871</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ</td>
<td>770</td>
<td>95</td>
<td>135</td>
<td>972</td>
<td>4</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ1</td>
<td>0</td>
<td>475</td>
<td>525</td>
<td>1</td>
<td>569</td>
<td>430</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ2</td>
<td>0</td>
<td>481</td>
<td>519</td>
<td>5</td>
<td>571</td>
<td>424</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SC</td>
<td>0</td>
<td>491</td>
<td>509</td>
<td>237</td>
<td>371</td>
<td>392</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GP</td>
<td>0</td>
<td>540</td>
<td>460</td>
<td>290</td>
<td>353</td>
<td>357</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>QIC</td>
<td>50</td>
<td>88</td>
<td>862</td>
<td>44</td>
<td>77</td>
<td>879</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ</td>
<td>646</td>
<td>148</td>
<td>206</td>
<td>934</td>
<td>5</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ1</td>
<td>0</td>
<td>445</td>
<td>555</td>
<td>0</td>
<td>535</td>
<td>465</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ2</td>
<td>0</td>
<td>443</td>
<td>557</td>
<td>10</td>
<td>535</td>
<td>455</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SC</td>
<td>0</td>
<td>467</td>
<td>533</td>
<td>168</td>
<td>397</td>
<td>435</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GP</td>
<td>0</td>
<td>549</td>
<td>451</td>
<td>269</td>
<td>406</td>
<td>325</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>QIC</td>
<td>16</td>
<td>39</td>
<td>945</td>
<td>7</td>
<td>33</td>
<td>960</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ</td>
<td>646</td>
<td>154</td>
<td>198</td>
<td>972</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ1</td>
<td>0</td>
<td>455</td>
<td>545</td>
<td>1</td>
<td>603</td>
<td>396</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RJ2</td>
<td>0</td>
<td>455</td>
<td>545</td>
<td>1</td>
<td>609</td>
<td>390</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SC</td>
<td>0</td>
<td>480</td>
<td>520</td>
<td>177</td>
<td>458</td>
<td>365</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GP</td>
<td>0</td>
<td>532</td>
<td>468</td>
<td>247</td>
<td>457</td>
<td>296</td>
</tr>
</tbody>
</table>

Conflicts of Interest

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The author was supported by a grant from the Penn State CTSI. The project was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant 5 UL1 RR0330184-04. The content is solely the responsibility of the author and does not represent the views of the NIH.

References


