Communications of the Korean Statistical Society
2011, Vol. 18, No. 6, 751–759

DOI: http://dx.doi.org/10.5351/CKSS.2011.18.6.751

Note on Working Correlation in the GEE of Longitudinal Counts Data

Kwang Mo Jeong[1,a]

[a]Department of Statistics, Pusan National University

Abstract

The method of generalized estimating equations(GEE) is widely used in the analysis of a correlated dataset that consists of repeatedly observed responses within subjects. The GEE uses a quasi-likelihood equations to find the parameter estimates without assuming a specific distribution for the correlated responses. In this paper we study the importance of specifying the working correlation structure appropriately in fitting GEE for correlated counts data. We investigate the empirical coverages of confidence intervals for the regression coefficients according to four kinds of working correlations where one structure should be specified by the users. The confidence intervals are computed based on the asymptotic normality and the sandwich variance estimator.

Keywords: Longitudinal counts data, GEE, working correlation structure, sandwich variance estimates.

1. Introduction

In clinical studies one may record the response variable repeatedly within each subject at several times or under various conditions. Repeated categorical responses commonly occur in biomedical applications of longitudinal studies. For example, a physician might observe patients conditions at weekly intervals regarding whether a new treatment is successful. Explanatory variables, usually called covariate variables, may also vary over time. Sometimes the responses refer to clusters of subjects. Repeated responses within a cluster tend to be more alike than observations from different clusters. Each subject may be regarded as a single cluster.

Ordinary analyses that ignore the correlations structure of repeatedly observed responses over time may be badly inappropriate. The GEE approach utilizes a covariance structure for the repeated responses without assuming any particular multivariate distribution. The alternative method for treating the longitudinal categorical responses is to use the generalized linear model(GLM) with random effects of subjects. The applications of GLM with random effects can be referred to Jeong (2005). The GEE is a multivariate version of quasi-likelihood that is computationally simpler than the GLM with random effects. The GEE provides consistent estimates when the model is correct in the sense that the link function and the linear predictor truly describe the model; however, the GEE is not a likelihood based approach and hence cannot use the methods of likelihood in testing fit, comparing models, and conducting inference about parameters. Inference on GEE parameters uses Wald statistics based on the asymptotic normality of the estimators together with their estimated covariance matrix. Firth (1993), and Kauermann and Caroll (2001) studied the asymptotic properties of standard errors in small sample sizes or in comparison to parametric estimator, respectively.

It is required to specify a working correlation structure when we fit GEE using common statistical packages; however, the true correlation structure of a given data set is unknown and we have

[1] Professor, Department of Statistics, Pusan National University, Pusan 609-735, Korea. E-mail: kmjung@pusan.ac.kr

no guideline in specifying the working correlation. In this paper we study the effects of working correlation to the asymptotic properties of GEE estimators based on the simulated counts having the assumed correlation structures. According to Liang and Zeger (1986) the independence working correlation can have surprisingly good efficiency when the correlation is week to moderate. There are other correlation structures which are commonly used in practice. The exchangeable correlation, the autoregressive of lag one(AR1) and the unstructured correlation structure. These correlation structures will be explained in Section 2.1. All working correlation structures yield similar GEE estimates and standard errors when the correlations are modest. For the case of clustered ordinal data Nores and Diez (2008) investigated some properties of GEE according to working correlation structures. They compared the coverage probability of confidence interval and the efficiency in the sense of variance estimates. The asymptotic efficiency of a correctly specified exchangeable association structure relative to the independence was discussed.

Through an empirical study we investigate the importance of specifying the working correlation appropriately which is close to the true covariance structure of a given dataset. In Chapter 2 we introduce the general framework of GEE in the respect of correlated counts responses by explaining commonly used working correlation structures. We briefly review the covariance matrix of repeated Poisson counts and the variance estimator of GEE regression coefficients. Through a practical example we illustrate the importance of properly specifying the working correlation in using statistical packages. Based on a Monte Carlo study we finally provide a suggestive guide in choosing the appropriate working correlation types. We expect the improved efficiency by choosing the working correlation wisely.

2. Generalized Estimating Equations Method

2.1. Framework of GEE

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iT})'$ be a multivariate response consisting of counts Y_{ij} repeatedly observed over T times for the i^{th} subject, where $i = 1, 2, \ldots, N$, and T sometimes varies by subject but we assume that T is fixed for every subject. The quasi-likelihood method assumes a model for $\mu_{ij} = E(Y_{ij})$ and specifies a variance function $v(\mu_{ij})$ describing how the variance $\text{var}(Y_{ij})$ depends on μ_{ij}. This method also requires a working guess for the correlation structure among $\{Y_{ij}\}$. To incorporate the covariate variables in a GEE model we denote the $p \times 1$ vector of explanatory variables by $\mathbf{x}_{ij} = (x_{1ij}, x_{2ij}, \ldots, x_{pij})'$ for the observed y_{ij}. The explanatory variables may vary for the repeated measurements.

We assume that the marginal means of responses are related to the explanatory variables $\mathbf{x}_{ij}$ in terms of link $g(\cdot)$ by the model

$$\eta_{ij} = g(\mu_{ij}) = \mathbf{x}_{ij}'\boldsymbol{\beta}. \tag{2.1}$$

In GEE we further assume that $\text{var}(Y_{ij}) = \phi v(\mu_{ij})$ for a known variance function $v(\mu_{ij})$ and a common scale parameter ϕ. The commonly used variance function for the Poisson counts is set by $v(\mu_{ij}) = \mu_{ij}$. If the scale parameter ϕ is greater than one we doubt overdispersion since $\text{var}(Y_{ij}) > \mu_{ij}$. In solving GEE a working correlation matrix $\mathbf{R}(\boldsymbol{\rho})$ for $\mathbf{Y}_i$ has an important role, that depends on a vector $\boldsymbol{\rho}$ of correlation parameters. The working covariance matrix for $\mathbf{Y}_i$ is given in terms of $\mathbf{R}(\boldsymbol{\rho})$ and variance function $v(\mu_{ij})$ as follows

$$\mathbf{V}_i = \phi \mathbf{A}_i^{\frac{1}{2}} \mathbf{R}(\boldsymbol{\rho}) \mathbf{A}_i^{\frac{1}{2}}, \tag{2.2}$$

where $\mathbf{A}_i = \text{diag}(v(\mu_{ij}))$. The working covariance matrix $\mathbf{V}_i$ in (2.2) coincides with the covariance matrix $\text{cov}(\mathbf{Y}_i)$ when $\mathbf{R}(\boldsymbol{\rho})$ is the true correlation matrix for $\mathbf{Y}_i$.

Let $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ be a $T\times p$ matrix with typical element $\partial\mu_{ij}/\partial\beta_k$, where $\mu_{ij} = g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta})$. According to the general discussion by Agresti (2002), the parameter estimators are obtained by solving the GEE given by

$$\sum_{i}^{N} \mathbf{D}_i'\mathbf{V}_i^{-1}\left[\mathbf{Y}_i - \boldsymbol{\mu}_i\right] = \mathbf{0}. \tag{2.3}$$

The GEE was firstly proposed by Liang and Zeger (1986) for marginal modeling with GLMs. When $\mathbf{R}(\boldsymbol{\rho})$ equals the identity matrix, the GEE treats pairs of responses as independent, and the working covariance matrix reduced to $\mathbf{V}_i = \phi\mathbf{A}_i$. In this case the GEE estimator $\hat{\boldsymbol{\beta}}$ is then the same as the ordinary estimator for a GLM treating the repeated responses as independent.

Instead of assuming the independence structure it would be more desirable to choose a working correlation structure permitting dependence between repeated responses. We may treat $\text{corr}(Y_{ij}, Y_{ik}) = \rho, \rho^{|k-j|}$ or ρ_{jk}. These types of correlation structures are called the exchangeable correlation, the AR1, and the unstructured correlation structure, respectively. The unstructured correlation structure seems to be more flexible and realistic than other correlation types; however, we should be cautious on the deficiencies due to additional parameters incurred by a separate correlation for each pair.

2.2. Sandwich variance estimator

Thall and Vail (1990) suggested some covariance models for longitudinal counts data with overdispersion by considering both the subject effects and the time effects. They derived the variances and covariances of Y_{ij} and Y_{ik} as

$$\begin{aligned}\sigma_{ij}^2 &= \text{var}\left(Y_{ij}\right) = \mu_{ij} + \left(\alpha_0 + \alpha_j\right)\mu_{ij}^2,\\ \sigma_{ijk} &= \text{cov}\left(Y_{ij}, Y_{ik}\right) = \alpha_0\mu_{ij}\mu_{ik},\end{aligned}$$

where α_0 is a subject variance scaled by the square of its mean and α_j is the variance effect mixed with both time effects and subject effects. There are several important simplifications and variants of this formulation, and we refer to Thall and Vail (1990) for detailed discussions. If the covariance terms are set to be $\sigma_{ijk} = \rho_{jk}\sigma_{ij}\sigma_{ik}$ for some suitable forms of σ_{ij} and correlation ρ_{jk}, this formulation coincides with the one proposed by Liang and Zeger (1986); however, allowing time-varying overdispersion. This formulation may be more desirable in settings where correlations that vary with (j,k) are appropriate. A more parsimonious approach is to take $\rho_{ijk} = \rho$ or $\rho^{|k-j|}$ as commented in Section 2.1.

Now we explain the consistent variance estimator of $\hat{\boldsymbol{\beta}}$. Let

$$\boldsymbol{\Sigma}_N = N\boldsymbol{\Sigma}_0\left[\sum_{i}^{N}\mathbf{D}_i'\mathbf{V}_i^{-1}\text{cov}(\mathbf{Y}_i)\mathbf{V}_i^{-1}\mathbf{D}_i\right]\boldsymbol{\Sigma}_0, \tag{2.4}$$

where $\boldsymbol{\Sigma}_0 = [\sum_{i=1}^{N}\mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i]^{-1}$. A consistent estimator of $\boldsymbol{\Sigma}_N$ is obtained by a sample analog replacing $\boldsymbol{\mu}_i$ by $\hat{\boldsymbol{\mu}}_i$ and the other unknown parameters $\boldsymbol{\beta}$, ϕ and $\boldsymbol{\rho}$ by their GEE estimates, and the $\text{cov}(\mathbf{Y}_i)$ by $(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)'$. This estimator of covariance matrix of $\hat{\boldsymbol{\beta}}$ is called a sandwich estimator, because the empirical evidence is sandwiched between the model-driven covariance matrix. The $\boldsymbol{\Sigma}_N/N$ simplifies

Table 1: Estimated parameter estimates and their standard errors

Assumed working correlation		α	β_1	β_2	β_3	β_4	β_{12}	ϕ	ρ
Indep	1)	−2.3802	0.9497	−1.3481	0.7855	−0.1565	0.5506	4.390	
	2)	0.8057	0.0964	0.4553	0.2330	0.0658	0.1847	1.110	-
	3)	0.0031	<2e-16	0.0031	0.0008	0.0175	0.0029	-	
Exch		−2.4017	0.9507	−1.3459	0.7906	−0.1565	0.5514	4.397	0.3642
		0.8183	0.0986	0.4590	0.2359	0.0658	0.1857	1.111	0.0647
		0.0033	<2e-16	0.0034	0.0008	0.0175	0.0030	-	-
AR1		−2.6455	0.9441	−1.5080	0.8714	−0.1498	0.6127	4.460	0.5130
		0.8209	0.0924	0.4470	0.2386	0.0926	0.1807	1.120	0.0636
		0.0013	<2e-16	0.0007	0.0003	0.1058	0.0007	-	-
Unstr		−2.5963	0.9383	−1.4983	0.8602	−0.1520	0.6111	4.440	
		0.8379	0.0928	0.4521	0.2421	0.0785	0.1819	1.130	-
		0.0020	<2e-16	0.0009	0.0004	0.0529	0.0009	-	

1) GEE estimate, 2) Standard Error (S. E.) of GEE estimate, 3) P-value

to $\mathbf{\Sigma}_0$ provided that the working correlation structure is the true one and $\text{cov}(\mathbf{Y}_i) = \mathbf{V}_i$. But the true variance function is unknown in practice.

The asymptotic normality of $\hat{\boldsymbol{\beta}}$ follows by the GEE theory, for example, of Liang and Zeger (1986). Under regularity conditions, $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$, where $\mathbf{\Sigma} = \lim_{N\to\infty} \mathbf{\Sigma}_N$ with $\mathbf{\Sigma}_N$ defined in (2.4). It is known that both the GEE estimator $\hat{\boldsymbol{\beta}}$ and its sandwiched covariance estimator are consistent even with incorrect specification of the variance function. However, some efficiency loss occurs when the chosen variance function $v(\boldsymbol{\mu}_i)$ is badly inaccurate or the number of subjects N is small.

2.3. An example of longitudinal counts data

As an illustration, we present an analysis for the dataset from Thall and Vail (1990), which consist of 59 epileptic patients suffering from simple or complex partial seizures. At each of four successive clinical visits, the number of seizures occurring during a 2-week period was recorded. The explanatory variables appearing in the model are baseline seizure rate (x_1), computed as the logarithm of 1/4 the 8-week seizure counts, binary indicator of treatment or placebo (x_2), the logarithm of age in years (x_3), and the indicator denoting fourth visit (x_4). The sample correlation matrix of four repeated measurements of seizure counts is

$$\begin{pmatrix} 1.0000 & 0.8708 & 0.7377 & 0.8930 \\ 0.8708 & 1.0000 & 0.8025 & 0.8945 \\ 0.7377 & 0.8025 & 1.0000 & 0.8242 \\ 0.8930 & 0.8945 & 0.8242 & 1.0000 \end{pmatrix}.$$

We assume the following GLM relationship

$$\log(\mu_{ij}) = \alpha + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \beta_{12} x_{1ij} * x_{2ij}.$$

The pairwise sample correlations seem to be approximately uniform and we may specify the exchangeable correlation structure in finding GEE estimators. Table 1 shows the results of GEE fitting obtained from four types of correlation specifications; independence (Indep), exchangeable (Exch), AR1, and unstructured (Unstr). The GEE estimates, the their sandwich standard errors, and the corresponding P-values using Wald test are listed for regression coefficients.

The P-values of some regression coefficients, among others those of β_2 are very different according to working correlation structures. In particular, the independence and the exchange correlation

represent similar P-values compared to those of AR1 and the unstructured. We see a similar pattern in the estimates of both β_3 and β_{12}. We need to be careful in specifying the working correlation appropriately to approximate the covariance structure of a given dataset.

3. A Monte Carlo Study

3.1. Design of experiment

We consider a relationship given by the log regression model of the form

$$\log(\mu_{ij}) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2ij}. \tag{3.1}$$

The covariate x_{1ij} is taken as -1 for $i \leq 1/4, 0$ for $1/4 < i \leq 3/4$, and 1 elsewhere. We generate the variable x_{2ij} from $U(0, 1)$, the uniform distribution over $(0, 1)$. We note that x_{2ij} is time dependent but x_{1ij} is not. We take the number of repeated times to be $T = 4$, and the number of subjects as $N = 60, 120$. We let $\alpha = 0$, and $\beta_1 = \beta_2 = 1$. The repeated measurements of responses are generated from the correlated Poisson distribution having the specified correlation structures such as exchangeable or AR1 correlation matrices. The marginal mean μ_{ij} of Poisson counts can be computed from (3.1). The number of repetitions is set to be 1,000.

The simulation study has been implemented through R software and library functions. We may refer to R Development Core Team (2006) for the R language and its environment. The library function rcounts.reg is useful in generating correlated responses having specified correlation structure and marginal means. Several packages to fit the GEE are available in R but we applied the geeglm in geepack to fit GEE, which is computationally intensive and can be freely available from CRAN site (http://cran.r-project.org). For multivariate data, geepack allows covariate in the mean, scale, and correlation structure by separate link functions that provides sandwich and versions of jackknife variance estimators for all parameter estimates. For a detailed discussion of geeglm refer to Yan (2002), and also to Halekoh *et al.* (2006) for the discussion of geepack. We computed the empirical coverages of confidence intervals for β_1 and β_2, and also their average lengths among 1000 iterations.

3.2. Results of simulation study

The empirical coverages of confidence intervals for β_1 and β_2 are listed in Table 2 and Table 3 according to the assumed working correlation structure, the sample sizes, the nominal confidence level, and the correlation parameter ρ. The empirical coverages do not attain the nominal confidence levels when $N = 60$ but it increases to the nominal levels as sample size and ρ increase. When the true correlation structure is exchangeable with $\rho = 0.5$, the lengths of confidence intervals under the independence working correlation are wider than others, and the unstructured correlation has the shortest lengths; however the coverages for β_1 and β_2 are unstable under the unstructured correlation.

The results of exchangeable correlation and AR1 are very similar when the true correlation structure of repeated responses is exchangeable with $\rho = 0.7$ or 0.9. As we see in Table 2 the lengths of confidence intervals for the exchangeable correlation structure are shorter than those under AR1 when the true correlation structure is correctly assumed as exchangeable. On the other hand the empirical coverages of AR1 are sometimes better than those of exchangeable working correlation specification. The shorter lengths of confidence intervals under exchangeable working correlation seem to be attained as an expense of slightly lower coverages. The results of Table 3 denote a reversed pattern when we assume the AR1 as a true correlation structure. This means that the standard errors of GEE estimators are smaller when we choose the appropriate working correlation structure that approximate the covariance structure of a given dataset.

Table 2: Empirical coverages of confidence limits for β_1 and β_2 among 1000 repetitions when the true correlation structure is exchangeable with $\rho = 0.5, 0.7, 0.9$

(a) $\rho = 0.5$

Assumed working correlation	N	Confidence levels 90% β_1	90% β_2	95% β_1	95% β_2	99% β_1	99% β_2
Indep	60	1) 0.8750	0.8540	0.9090	0.8990	0.9770	0.9780
		2) 0.1930	0.2415	0.2300	0.2878	0.3028	0.3788
	120	0.8870	0.8940	0.9290	0.9330	0.9920	0.9890
		0.1334	0.1679	0.1589	0.2001	0.2092	0.2634
Exch	60	0.8700	0.8540	0.9120	0.8900	0.9730	0.9640
		0.1912	0.2335	0.2279	0.2782	0.2999	0.3663
	120	0.8850	0.8980	0.9360	0.9420	0.9920	0.9850
		0.1324	0.1637	0.1578	0.1950	0.2077	0.2567
AR1	60	0.8750	0.8490	0.9040	0.8860	0.9730	0.9730
		0.1903	0.2350	0.2267	0.2800	0.2984	0.3686
	120	0.8730	0.8870	0.9320	0.9310	0.9880	0.9900
		0.1314	0.1639	0.1566	0.1953	0.2061	0.2570
Unstr	60	0.8960	0.8250	0.9370	0.9070	0.9780	0.9690
		0.1612	0.2015	0.1921	0.2401	0.2529	0.3160
	120	0.8720	0.8800	0.9400	0.9300	0.9810	0.9790
		0.1151	0.1438	0.1372	0.1713	0.1805	0.2255

(b) $\rho = 0.7$

Assumed working correlation	N	90% β_1	90% β_2	95% β_1	95% β_2	99% β_1	99% β_2
Indep	60	0.8780	0.8650	0.9290	0.9030	0.9730	0.9690
		0.3081	0.3361	0.3671	0.4005	0.4832	0.5271
	120	0.8910	0.8980	0.9590	0.9540	0.9930	0.9870
		0.2184	0.2398	0.2603	0.2857	0.3426	0.3760
Exch	60	0.8920	0.8730	0.9390	0.9350	0.9680	0.9790
		0.3006	0.2881	0.3581	0.3433	0.4714	0.4519
	120	0.9050	0.9030	0.9470	0.9470	0.9930	0.9800
		0.2126	0.2030	0.2534	0.2419	0.3335	0.3184
AR1	60	0.8940	0.8670	0.9530	0.9150	0.9770	0.9790
		0.3150	0.3158	0.3753	0.3763	0.4941	0.4953
	120	0.8990	0.9010	0.9530	0.9330	0.9930	0.9730
		0.2235	0.2238	0.2662	0.2666	0.3505	0.3509
Unstr	60	0.8840	0.8500	0.9250	0.9000	0.9840	0.9640
		0.2717	0.2515	0.3237	0.2996	0.4261	0.3944
	120	0.8750	0.9150	0.9320	0.9610	0.9860	0.9730
		0.1927	0.1751	0.2296	0.2087	0.3022	0.2747

(c) $\rho = 0.9$

Assumed working correlation	N	90% β_1	90% β_2	95% β_1	95% β_2	99% β_1	99% β_2
Indep	60	0.8680	0.8620	0.9110	0.9130	0.9850	0.9810
		0.4107	0.4308	0.4894	0.5133	0.6441	0.6756
	120	0.9260	0.8550	0.9690	0.9350	0.9940	0.9870
		0.2917	0.3069	0.3476	0.3656	0.4575	0.4813
Exch	60	0.8630	0.8750	0.9180	0.9160	0.9760	0.9750
		0.3463	0.1551	0.4127	0.1848	0.5432	0.2432
	120	0.9360	0.8760	0.9490	0.9360	0.9880	0.9940
		0.2450	0.1044	0.2920	0.1244	0.3843	0.1637
AR1	60	0.8780	0.8840	0.9250	0.9200	0.9670	0.9820
		0.4172	0.1821	0.4971	0.2169	0.6543	0.2855
	120	0.8940	0.8770	0.9550	0.9350	1.0000	0.9820
		0.2978	0.1233	0.3548	0.1469	0.4671	0.1934
Unstr	60	0.7750	0.7891	0.8375	0.8315	0.8799	0.8708
		1.4660	1.0356	1.7472	1.2340	2.2998	1.6243
	120	0.8450	0.8250	0.8900	0.8700	0.9360	0.9420
		0.4729	0.2966	0.5635	0.3533	0.7417	0.4651

1) Empirical coverage of confidence interval, 2) Length of confidence interval

Table 3: Empirical coverages of confidence limits for β_1 and β_2 among 1000 repetitions when the true correlation structure is AR1 with $\rho = 0.5, 0.7, 0.9$

(a) $\rho = 0.5$

Assumed working correlation	N	Confidence levels 90% β_1	90% β_2	95% β_1	95% β_2	99% β_1	99% β_2
Indep	60	0.8900	0.8840	0.9320	0.9340	0.9950	0.9900
		0.1626	0.2210	0.1938	0.2634	0.2551	0.3467
	120	0.9100	0.9140	0.9300	0.9740	0.9880	0.9940
		0.1107	0.1519	0.1318	0.1810	0.1736	0.2382
Exch	60	0.8810	0.9090	0.9370	0.9580	0.9800	0.9900
		0.1603	0.2075	0.1910	0.2473	0.2515	0.3255
	120	0.8630	0.8780	0.9370	0.9540	0.9810	0.9940
		0.1090	0.1412	0.1299	0.1682	0.1710	0.2215
AR1	60	0.8750	0.8960	0.9170	0.9290	0.9900	0.9950
		0.1588	0.2204	0.1892	0.2627	0.2490	0.3457
	120	0.9180	0.9210	0.9430	0.9810	0.9880	0.9940
		0.1081	0.1520	0.1287	0.1811	0.1695	0.2384
Unstr	60	0.8910	0.9100	0.9320	0.9370	0.9660	0.9810
		0.1564	0.2406	0.1864	0.2866	0.2453	0.3773
	120	0.8890	0.8970	0.9500	0.9360	0.9750	0.9810
		0.0893	0.1149	0.1064	0.1369	0.1400	0.1803

(b) $\rho = 0.7$

Assumed working correlation	N	Confidence levels 90% β_1	90% β_2	95% β_1	95% β_2	99% β_1	99% β_2
Indep	60	0.8780	0.8460	0.9160	0.9230	0.9710	0.9720
		0.2366	0.2773	0.2819	0.3304	0.3711	0.4350
	120	0.8990	0.8830	0.9510	0.9400	0.9880	0.9870
		0.1676	0.1966	0.1997	0.2342	0.2628	0.3083
Exch	60	0.8780	0.8670	0.9210	0.9140	0.9710	0.9720
		0.2355	0.2703	0.2806	0.3221	0.3693	0.4239
	120	0.9180	0.8900	0.9560	0.9470	0.9820	0.9870
		0.1668	0.1929	0.1987	0.2298	0.2616	0.3025
AR1	60	0.8830	0.8750	0.9160	0.9130	0.9710	0.9760
		0.2303	0.2614	0.2744	0.3115	0.3612	0.4100
	120	0.9190	0.9030	0.9560	0.9290	0.9820	0.9930
		0.1629	0.1848	0.1941	0.2201	0.2555	0.2898
Unstr	60	0.8630	0.8630	0.9240	0.8800	0.9730	0.9730
		0.2190	0.2190	0.2609	0.2983	0.3435	0.3926
	120	0.9050	0.8900	0.9500	0.9410	0.9820	0.9870
		0.1540	0.1711	0.1835	0.2038	0.2415	0.2683

(c) $\rho = 0.9$

Assumed working correlation	N	Confidence levels 90% β_1	90% β_2	95% β_1	95% β_2	99% β_1	99% β_2
Indep	60	0.8610	0.8570	0.9200	0.9210	0.9810	0.9680
		0.3827	0.4017	0.4559	0.4786	0.6002	0.6300
	120	0.9250	0.8980	0.9790	0.9360	0.9930	0.9870
		0.2717	0.2878	0.3237	0.3429	0.4261	0.4514
Exch	60	0.8790	0.8950	0.9450	0.9170	0.9860	0.9590
		0.3507	0.2316	0.4179	0.2760	0.5501	0.3633
	120	0.9190	0.8970	0.9790	0.9360	0.9860	0.9930
		0.2489	0.1610	0.2965	0.1918	0.3903	0.2525
AR1	60	0.8760	0.8850	0.9360	0.9260	0.9900	0.9670
		0.3492	0.2170	0.4161	0.2585	0.5477	0.3403
	120	0.9520	0.9010	0.9660	0.9730	0.9930	0.9860
		0.2472	0.1486	0.2945	0.1770	0.3877	0.2330
Unstr	60*	0.8170	0.8580	0.8760	0.9000	0.9080	0.9350
		2.9010	2.5120	3.4570	2.9930	4.5510	3.9390
	120	0.8990	0.9060	0.9450	0.9720	0.9860	0.9930
		0.3245	0.1524	0.3866	0.1815	0.5089	0.2390

$*$ Unavailable estimates of β_1 and β_2 in a frequency of 9 cases among 1000 repetitions

We also comment that the lengths of confidence intervals, and hence the standard errors of $\hat{\beta}_k$, $k = 1, 2$, are smallest under the unstructured correlation structure but the empirical coverages are very unstable when $N = 60$. There have been unavailable GEE estimates in a frequency of about nine cases among 1000 repetitions under the unstructured working correlation under the AR1 with $N = 60$ and $\rho = 0.9$. We should be cautious on specifying the unstructured correlation structure because the performance is not good in many cases with respect to the empirical coverages and the length of confidence intervals. Furthermore there sometimes occurred no convergence of GEE estimates when sample size is small to moderate and the correlation parameter ρ is large.

4. Summary and Further Researches

In this study we investigated the effect of working correlation structure to the GEE estimates and their sandwich standard errors in terms of empirical coverages of confidence interval in GEE of longitudinal counts dataset. Four kinds of correlation structures; the independence, the exchangeable, the AR1 and the unstructured, are available in fitting GEE using the statistical packages such as R and SAS. We explained the difference in coverages of confidence intervals between four kinds of specifications when the true correlation structures are assumed as exchangeable or AR1 with several correlation parameters. From a small scale Monte Carlo study we found that the coverages and the lengths of confidence intervals depend on the choosing of working correlation structure. The specifications of exchangeable working correlation and AR1 are good in many cases in the respect of shorter lengths of confidence intervals but sometimes they have slightly lower coverages compared to the independence working correlation. The GEE estimates and their standard errors seems to be proper when we correctly specify the working correlation structure which closely approximate the true correlation matrix of a given dataset. We need to be careful to the specification of working correlation in fitting GEE.

In particular, we should be cautious on choosing the unstructured correlation structure among others because the GEE estimates are sometimes unavailable or greatly unstable when the sample size is small and the correlation parameter is large. As a further work it would be interesting to study the sensitivity of the correlation structure to the goodness-of-fit of GEE for the analysis of longitudinal counts data.

References

Agresti, A. (2002). *Categorical Data Analysis*, Second Edition, Wiley, New York.

Firth, D. (1993). Recent development in quasi-likelihood methods, *Proceedings of ISI 49th Session*, 341–358.

Halekoh, U., Hojsgaard, S. and Yan, J. (2006). The R package geepack for generalized estimating equations, *Journal of Statistical Software*, **15**, 1–11.

Jeong, K. M. (2005). Generalized linear mixed models for ordinal response in clustered data, *Journal of the Korean Data Analysis Society*, **7**, 817–828.

Kauermann, G. and Caroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation, *Journal of the American Statistical Association*, **96**, 1387–1397.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data using generalized linear models, *Biometrika*, **73**, 13–22.

Nores, M. L. and Diez, M. P. (2008). Some properties of regression estimates in GEE models for clustered ordinal data, *Computational Statistics & Data Analysis*, **52**, 3877–3888.

R Development Core Team (2006). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL: http://www.r-

project.org.
Thall, P. F. and Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion, *Biometrics*, **46**, 657–671.
Yan, J. (2002). Geepack: Yet another package for generalized estimating equations, *R News*, **2**, 12–14.

Received October 2011; Accepted October 2011