A bias-correction in Rotnitzky-Jewell criteria for improving the approach of correlation structure selection in generalized estimating equations

Ajmery Jaman*
James P Grant School of Public Health, Dhaka-1212, Bangladesh - ajaman@isrt.ac.bd

Abstract

In generalized estimating equations (GEE), the correlation between the repeated observations on a subject is modeled with a patterned working correlation matrix. Specifying the working correlation structure correctly is gainful, in terms of improving efficiency and enhancing scientific understanding. For analyzing cluster correlated data such as longitudinal data, a number of criteria are available in the literature for selecting an appropriate working correlation structure in GEE. The Rotnitzky-Jewell (RJ) criteria, which we have considered in this paper based on their good performance, are based on the fact that if the assumed working correlation structure is correct then the model-based (naive) and the sandwich (robust) covariance estimators of the regression coefficient estimates should be close to each other. In this paper, we propose a set of new criteria modifying the RJ criteria based on the bias-corrected sandwich covariance estimator, and show a comparison between the proposed criteria and the RJ criteria via a simulation study using correlated binary response. The results revealed that the proposed bias correction approach brings improvement in the RJ criteria in terms of improving the percentage selection of the correct correlation structure.

Keywords: bias-corrected sandwich covariance estimator; longitudinal data; model-based covariance estimator; working correlation structure.

1. Introduction

Generalized estimating equations (GEE) offer a regression methodology for cluster correlated data and provide a marginal analysis of the data that accounts for the correlation between the responses within the same cluster (Liang and Zeger, 1986). GEE give consistent estimators of the regression coefficients and of their variances under weak assumptions about the actual correlation among the observations for each subject. But the efficiency of the estimators depends on the correct specification of the working correlation structure, and a working correlation structure that closely approximates the true underlying structure results in better precision (Pepe and Anderson, 1994). A number of criteria are available in the literature for selecting appropriate working correlation structure in GEE. Shults et al. (2009) studied the Rotnitzky-Jewell (RJ) criteria (Rotnitzky and Jewell, 1990) which are based on the comparison between the model-based covariance matrix and the sandwich-based covariance matrix of the regression parameter estimates. Shults et al. (2009) also compared the RJ criteria with the Shults-Chaganty (SC) criterion (Shults and Chaganty, 1998) which is based on the weighted error sum of squares. They showed that the SC criterion has relatively poor performance for correct selection of the working correlation structure in GEE. However, the RJ criteria involve the sandwich covariance estimator which is biased downward (Mancl and DeRouen, 2001) and generally has a larger variability than the model-based covariance estimator (Kauermann and Carroll, 2001). Use of the negatively-biased covariance estimator can result in the hypothesis tests of the regression coefficients that are too liberal and confidence intervals on the regression parameters that are too narrow, especially in smaller samples. To address the problem, few alternative approaches are available in the literature that estimate the covariance matrix of the marginal model parameter estimates more precisely by removing the inherent bias. In this article, we propose a simple modification to the RJ criteria replacing the robust sandwich covariance estimator in it with a bias-corrected sandwich covariance estimator which was proposed by Mancl and DeRouen (2001).

In the next section, we give a brief description of the GEE approach to introduce the notations used in this article. A brief description of the RJ criteria is provided in Section 2.2. The definition of the proposed criteria is given in Section 2.3. In Section 3, we present the results from a simulation study that evaluates the performance of the proposed criteria and compares the criteria with the RJ criteria in terms of the selection

of true underlying structure for correlated binary responses in marginal regression models. Finally a general discussion is provided in Section 4.

2. Criteria for Correlation Structure Selection in GEE

In this section, we introduce some notation with a brief discussion on GEE, briefly review the existing criteria, and propose the new set of criteria for appropriate correlation structure selection in GEE.

2.1 Notations

Let y_{ij} be the response and $\mathbf{x}_{ij}$ be the $p \times 1$ vector of covariates at jth time for ith subject ($i = 1, 2, \ldots, N$, and $j = 1, 2, \ldots, n_i$). Let $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in_i})'$ is the $n_i \times 1$ vector of responses and $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{in_i})'$ is the $n_i \times p$ matrix of covariates for subject i. Response vectors for two different subjects (say $\mathbf{y}_i$ and $\mathbf{y}_{i^\partial}$) are assumed to be independent, but generally responses are correlated within each subject. The marginal distribution that requires specifying marginal mean and variance of y_{ij} is specified by a generalized linear model (GLM) (McCullagh and Nelder, 1989) as: i) $g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}$, where $\mu_{ij} = E(Y_{ij}|\mathbf{x}_{ij})$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ is the regression coefficient vector and g is the link function, and ii) $var(Y_{ij}) = \phi v(\mu_{ij})$, where v is a known function and ϕ is a scale parameter which may need to be estimated. Estimation of the unknown regression coefficient vector $\boldsymbol{\beta}$ is usually the interest. The covariance matrix of $\mathbf{y}_i$ is specified as $\mathbf{V}_i = \mathbf{A}_i^{1=2}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{1=2}$, where $\mathbf{A}_i$ is a $n_i \times n_i$ diagonal matrix with elements $var(Y_{ij}) = \phi v(\mu_{ij})$ as the jth diagonal element, $\mathbf{R}_i(\boldsymbol{\alpha})$ is the correlation matrix among the outcomes measured at different times for the ith subject and $\boldsymbol{\alpha}$ is a vector of unknown parameters that completely specifies within subject correlation. The GEE approach estimates $\boldsymbol{\beta}$ by solving the estimating equations (Liang and Zeger, 1986)

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbf{S}_i(\boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbf{D}'_i \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}\,, \tag{1}$$

where $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}'$. $\mathbf{R}_i(\boldsymbol{\alpha})$ is known as the "working" correlation matrix, and to solve the Generalized Estimating Equations (1) we must define the structure of $\mathbf{R}_i(\boldsymbol{\alpha})$. The choices of working correlation structure may include exchangeable or Compound Symmetry (CS), the first order autoregressive (AR(1)), toeplitz, and unstructured correlation structure.

Liang and Zeger (1986) proposed that when the marginal mean $\boldsymbol{\mu}_{ij}$ is correctly specified and when mild regularity conditions hold, the GEE estimate $\hat{\boldsymbol{\beta}}$ obtained under an assumed working correlation structure is asymptotically multivariate normal with mean vector $\boldsymbol{\beta}$ and covariance matrix

$$V_S = B^{-1}\Big[\sum_{i=1}^{N} \mathbf{D}'_i \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}\mathbf{D}_i\Big]B^{-1}\,, \tag{2}$$

where $B = \sum_{i=1}^{N} D'_i V_i^{-1} D_i$. The matrix V_S is known as the sandwich covariance matrix and the corresponding empirical version is known as the sandwich covariance estimator. However, if the modeling specifications (i) and (ii) are correct then there is no difference between the sandwich estimator and the model-based estimator, $V_M = (\sum_{i=1}^{N} \mathbf{D}'_i \mathbf{V}_i^{-1}\mathbf{D}_i)^{-1}$. The estimation of the correlation parameter vector $\boldsymbol{\alpha}$ is usually done using the method (typically known as the second-order GEE or simply GEE2) proposed by Prentice and Zhao (1991), which uses a second set of estimating equations for $\boldsymbol{\alpha}$. GEE2 is less robust to misidentification of the correlation structure than is usual moment based estimation of the correlation parameters (Shults and Hilbe, 2014, p. 59).

2.2 RJ Criteria

The Rotnitzky-Jewell (RJ) criteria (Rotnitzky and Jewell, 1990) compare the model-based estimators $\hat{V}_M$ (that assumes correct specification) and the sandwich-based estimators $\hat{V}_S$ (that is typically robust to misidentification of the correlation structure) of the covariance matrix of $\hat{\boldsymbol{\beta}}$ under the assumed working correlation structure. If the working correlation structure is close to the true structure, the model-based and sandwich-based estimates of the covariance matrix should be similar, and consequently, both $Q = \hat{V}_M^{-1}\hat{V}_S$ and Q^2

should be close to a $p \times p$ identity matrix. Following this the RJ criteria are defined as

$$\begin{aligned} \text{RJ1} &= trace(Q)/p \\ \text{RJ2} &= trace(Q^2)/p \\ \text{DBAR} &= \sum_j (e_j - 1)^2 = \text{RJ2} - 2\text{RJ1} + 1\,, \end{aligned} \tag{3}$$

where the e_j are the eigenvalues of Q and p is the dimension of the parameter vector $\boldsymbol{\beta}$. For RJ1 the structure corresponding to the minimum value of $|\text{RJ1} - 1|$ is selected; the same rule is applied for RJ2. For DBAR the structure corresponding to the minimum absolute value of DBAR is selected.

2.3 Proposed Bias-Correction in RJ Criteria

The RJ criteria involve the sandwich covariance estimator $\hat{V}_{\mathsf{S}}$. Due to the negatively biased nature of $\hat{V}_{\mathsf{S}}$, alternative variance-covariance estimators have been proposed by Mancl and DeRouen (2001) and Kauermann and Carroll (2001) by correcting the inherent bias in it. In this article, we are considering the bias corrected sandwich covariance estimator proposed by Mancl and DeRouen (2001).

In practice, to calculate the GEE robust covariance estimator, (2) the residual estimates, $\hat{\mathsf{r}}_i = \mathsf{y}_i - \hat{\boldsymbol{\mu}}_i$, are used to estimate $cov(\mathsf{Y}_i)$. Using a first-order Taylor series expansion of the vector, $\hat{\mathsf{r}}_i$, about $\boldsymbol{\beta}$ it can be shown that

$$E\big[\hat{\mathsf{r}}_i \hat{\mathsf{r}}_i'\big] \approx (\mathsf{I}_{n_i} - \mathsf{H}_{ii}) cov[\mathsf{y}_i](\mathsf{I}_{n_i} - \mathsf{H}_{ii})' + \sum_{m \neq i} \mathsf{H}_{im} cov[\mathsf{y}_m] \mathsf{H}_{im}'\,, \tag{4}$$

where $\mathsf{H}_{im} = \mathsf{D}_i (\sum_{l=1}^{N} \mathsf{D}_l' \mathsf{V}_l^{-1} \mathsf{D}_l)^{-1} \mathsf{D}_m' \mathsf{V}_m^{-1}$ and I_{n_i} is an identity matrix of the same dimension as H_{ii}. In order to derive a tractable approximation to the bias, Mancl and DeRouen (2001) assumed that the contribution to the bias of the sum in expression (4) is negligible. By definition, the elements of H_{im} are between zero and one, usually close to zero, so it may be reasonable to assume that the summation makes only a small contribution to the bias. However, assuming that $E[\hat{\mathsf{r}}_i \hat{\mathsf{r}}_i']$ is approximated by $(\mathsf{I}_{n_i} - \mathsf{H}_{ii}) cov[\mathsf{y}_i](\mathsf{I}_{n_i} - \mathsf{H}_{ii})'$, the bias-corrected sandwich covariance estimator becomes

$$\hat{V}_{\mathsf{S}_{bc}} = B^{-1} \Big[\sum_{i=1}^{N} D_i' V_i^{-1} (I_{n_i} - H_{ii})^{-1} \hat{\mathsf{r}}_i \hat{\mathsf{r}}_i' (I_{n_i} - H_{ii})^{-1} V_i^{-1} D_i \Big] B^{-1}\,. \tag{5}$$

To improve the performance of the RJ criteria we replace the usual sandwich estimator $\hat{V}_{\mathsf{S}}$ by the bias corrected sandwich estimator $\hat{V}_{\mathsf{S}_{bc}}$ in (3) and propose the following modified criteria

$$\begin{aligned} \text{RJ1bc} &= trace(Q_{bc})/p \\ \text{RJ2bc} &= trace(Q_{bc}^2)/p \\ \text{DBARbc} &= \text{RJ2bc} - 2\text{RJ1bc} + 1\,, \end{aligned}$$

where $Q_{bc} = \hat{V}_M^{-1} \hat{V}_{\mathsf{S}_{bc}}$. The RJ1bc criterion chooses the structure that corresponds to the minimum value of $|\text{RJ1bc} - 1|$; the same rule is applied for RJ2bc criterion; and finally the DBARbc criterion chooses the structure corresponding to the minimum absolute value of DBARbc. The performances of the proposed criteria are evaluated and compared with the existing criteria by a simulation study with different simulation scenarios which is given in the next section.

3. Simulation Study

Our objectives are to compare the performance of the proposed criteria with the performance of RJ criteria in detecting the correct working correlation structure in GEE. For this comparison we conducted a simulation study with different simulation scenarios using correlated binary response. For data generation we consider that the marginal mean μ_{ij} of the response y_{ij} corresponding to jth time point in the ith cluster is specified as a function of covariates as

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij}\,, \quad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, n_i\,, \tag{6}$$

where both the covariates x_{1ij} and x_{2ij} are binary and generated from Bernoulli(0.5). The number of observations within the ith cluster, n_i, is fixed at 5 and same for all $i = 1, 2, \ldots, N$. The covariate x_1 is cluster-level, i.e., it takes the same values over all the observations within the same cluster, and the covariate x_2 is observation-level, i.e., it may take different values over the observations within the same cluster. For model (6) the true values of the regression parameters are fixed at $\beta_0 = -1.6$, $\beta_1 = 0.38$, and $\beta_2 = 0.35$. The correlation structure selection criteria are being compared for the situation with $\alpha = 0.3$ and $N = 50$.
For generating data with unstructured correlation structure, the correlation between observation j and j' within the same cluster i is defined as $\alpha^{|t_{ij} - t_{ij'}|^\lambda}$, where t indicates the time, which reduces to exchangeable and AR(1) structures for $\lambda = 0$ and $\lambda = 1$, respectively. For simulations $\lambda = 0.5$ and $t \in \{1, 2, 3, 5, 10\}$ are used. For toeplitz correlation structure the first $(n_i - 1)$ elements of $\alpha^{|t_{ij} - t_{ij'}|^\lambda}$ are used as the correlation parameter vector for the ith cluster. We compare the performances of the existing criteria with the performance of our proposed criteria in correct correlation structure selection in GEE in two settings:

1. For each criterion, the best approximating working correlation structure is chosen among independence, exchangeable or Compound Symmetry (CS) and AR-1 structures only.
2. For each criterion, the best approximating working correlation structure is chosen among independence, exchangeable, AR-1, toeplitz and unstructured correlation structures.

All computations are performed using *R* version 2.15.2, with GEE fitting performed using *geepack* library. Correlated binary response generation was performed using *binarySimCLF* library which implements the method described by Qaqish (2003).
Table 1 presents the percentage selections (out of 2000 independent replications) of true correlation structure alongside the selectins of other candidate correlation structures by the RJ and the corresponding proposed criteria for different simulation scenarios. Table 1 shows that for setting 1, when the true Intra-cluster Correlation Structure (ICS) is independence, the use of bias-corrected sandwich covariance estimator increases the correct detection of independence structure 7 percent of the time for RJ1 criterion, 8 percent of the time for RJ2 criterion and 15 percent of the time for DBAR criterion. But still in case of true independence structure the proposed criteria incorrectly select either CS or AR(1) structure more often than the true structure. On the other hand when the true ICS is exchangeable (or CS), bias correction to the RJ criteria brings 14 percent improvement to RJ1 criterion, 8 percent improvement in RJ2 criterion and 6 percent improvement to DBAR criterion. Moreover, the percentage selection of the true CS structure for RJ2bc criterion is 98. It implies that if the true correlation structure is exchangeable and we use RJ2bc criterion instead of using RJ2 the probability of making the wrong selection of working correlation structure given that the competing set includes independence, exchangeable and AR(1) structure is nearly zero. Further, when the true ICS is autoregressive of order 1 (AR(1)), the bias correction approach brings 12 percent improvement to each of the RJ1 and RJ2 criteria and 10 percent improvement to DBAR criterion.
For setting 2 with independent observations, the improvements due to bias correction approach are 7, 10 and 13 percent for RJ1, RJ2 and DBAR criteria, respectively. When the true correlation structure is CS, the correct selections of CS structure are between 32 and 36 percent for RJ criteria and between 36 and 38 percent for the proposed criteria indicating a little improvement due to bias correction. Further, if the true intra-cluster correlation structure is AR(1), the use of bias corrected approach brings more than ten percent improvement to the existing criteria. But, if we have data with the toeplitz or the unstructured correlation structure as the true ICS, none of the existing criteria performs well in identifying the true correlation structure, and also the bias correction doesn't bring any improvement to the original set of criteria.

4. Conclusions

The objectives of this article were to propose a simple modification to the RJ criteria to improve its performance regarding the choice of an appropriate working correlation structure in GEE and to compare the performance of the RJ and the proposed criteria in selecting the correct working correlation structure in GEE analysis of correlated binary response. The RJ criteria involve sandwich covariance estimator which is biased downward. Review of this topic motivated us to use the bias-corrected sandwich covariance estimator in RJ criteria and thus propose a modified set of criteria for appropriate working correlation structure selection in GEE. To compare the proposed criteria with the original set of RJ criteria we conducted a simulation study

Table 1: Percentage selection of working correlation structures by the competing selection criteria for different true intra-cluster correlation structures with true $\alpha = 0.3$ and $N = 50$ from 2000 independent replications.

True ICS	Criterion	Working Correlation Structures							
		Setting 1			Setting 2				
		Indep	CS	AR(1)	Indep	CS	AR(1)	Toep	UN
Indep	RJ1	**17**	61	22	**14**	23	18	19	26
	RJ1bc	**24**	51	25	**21**	19	23	18	20
	RJ2	**19**	60	21	**15**	24	17	20	25
	RJ2bc	**27**	45	28	**25**	15	25	15	20
	DBAR	**15**	64	21	**5**	21	8	23	43
	DBARbc	**29**	39	32	**18**	12	23	14	34
CS	RJ1	0	**82**	18	0	**32**	14	24	30
	RJ1bc	0	**96**	4	0	**38**	2	28	31
	RJ2	0	**90**	10	0	**35**	7	27	31
	RJ2bc	0	**98**	2	0	**38**	1	31	30
	DBAR	0	**89**	11	0	**36**	7	26	32
	DBARbc	0	**95**	5	0	**36**	2	29	33
AR(1)	RJ1	4	62	**34**	4	32	**21**	22	22
	RJ1bc	1	53	**46**	0	28	**34**	18	20
	RJ2	2	61	**37**	1	30	**23**	22	24
	RJ2bc	1	50	**49**	0	26	**37**	16	21
	DBAR	0	60	**40**	0	31	**13**	21	34
	DBARbc	0	49	**51**	0	25	**28**	17	30
Toep	RJ1				1	30	22	**21**	25
	RJ1bc				0	31	18	**22**	28
	RJ2				0	31	19	**23**	27
	RJ2bc				0	32	18	**22**	28
	DBAR				0	34	11	**21**	34
	DBARbc				0	31	15	**22**	32
UN	RJ1				3	27	22	22	**26**
	RJ1bc				1	29	19	24	**27**
	RJ2				2	29	20	24	**26**
	RJ2bc				1	28	19	24	**28**
	DBAR				0	30	9	21	**40**
	DBARbc				0	25	14	23	**37**

with correlated binary response. The simulation results revealed that, when the true correlation structure is either of the independence, exchangeable or AR(1) structure the proposed bias-corrected approach brings improvement to each of the corresponding RJ criteria in terms of improving the percentage selection of the correct correlation structure. But for the other two correlation structures (toeplitz and unstructured), performances of the RJ and the corresponding bias-corrected criteria are almost equal.

Acknowledgement

I want to acknowledge A. H. M. Mahbub Latif, Professor, Institute of Statistical Research and Training, University of Dhaka; Wasimul Bari, Professor, Department of Statistics, Biostatistics and Informatics, University

of Dhaka; and Abdus S. Wahed, Associate Professor, Department of Biostatistics, University of Pittsburgh for their contribution in this article.

References

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of longitudinal data*. Oxford University Press.

Kauermann, G. and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Mancl, L. A. and DeRouen, T. A. (2001). A covariance estimator for gee with improved small-sample properties. *Biometrics*, 57(1):126–134.

McCullagh, P. and Nelder, J. (1989). *General linear models*.

Pepe, M. S. and Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics-Simulation and Computation*, 23(4):939–951.

Prentice, R. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44(4):1033–1048.

Prentice, R. and Zhao, L. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47(3):825–839.

Qaqish, B. F. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, 90(2):455–463.

Rotnitzky, A. and Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77(3):485–497.

Shults, J. and Chaganty, N. R. (1998). Analysis of serially correlated data using quasi-least squares. *Biometrics*, 54(4):1622–1630.

Shults, J. and Hilbe, J. M. (2014). *Quasi-Least Squares Regression*. CRC Press.

Shults, J., Sun, W., Tu, X., Kim, H., Amsterdam, J., Hilbe, J. M., and Ten-Have, T. (2009). A comparison of several approaches for choosing between working correlation structures in generalized estimating equation analysis of longitudinal binary data. *Statistics in medicine*, 28(18):2338–2355.

Sommer, A. et al. (1982). *Nutritional blindness. Xerophthalmia and keratomalacia*. Oxford University Press.

Sommer, A., Katz, J., and Tarwotjo, I. (1984). Increased risk of respiratory disease and diarrhea in children with preexisting mild vitamin a deficiency. *The American journal of clinical nutrition*, 40(5):1090–1095.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American statistical association*, 86(413):79–86.