Statistica Sinica **12**(2002), 475-490

SELECTING THE WORKING CORRELATION STRUCTURE IN GENERALIZED ESTIMATING EQUATIONS WITH APPLICATION TO THE LUNG HEALTH STUDY

Wei Pan and John E. Connett

University of Minnesota

Abstract: The generalized estimating equation (GEE) approach is becoming more and more popular in handling correlated response data, for example in longitudinal studies. An attractive property of the GEE is that one can use some *working* correlation structure that may be wrong, but the resulting regression coefficient estimate is still consistent and asymptotically normal. One convenient choice is the independence model: treat the correlated responses as if they were independent. However with time-varying covariates there is a dilemma: using the independence model may be very inefficient (Fitzmaurice (1995)); using a non-diagonal working correlation matrix may violate an important assumption in GEE, producing biased estimates (Pepe and Anderson (1994)). It would be desirable to be able to distinguish these two situations based on the data at hand. More generally, selecting an appropriate working correlation structure, as an aspect of model selection, may improve estimation efficiency. In this paper we propose some resampling-based methods (i.e., the bootstrap and cross-validation) to do this. The methodology is demonstrated by application to the Lung Health Study (LHS) data to investigate the effects of smoking cessation on lung function and on the symptom of chronic cough. In addition, Pepe and Anderson's result is verified using the LHS data.

Key words and phrases: Bootstrap, cross-validation, GEE, GLM, model selection, PMSE.

1. Introduction

Correlated responses are common in biomedical studies. One typical example is the longitudinal study where each subject is followed over a period of time, and repeated observations of the response variable and relevant covariates are recorded. Since repeated observations are made on the same subject, observed responses are generally correlated. For continuous responses that can be treated as approximately normal, the linear mixed-effects models can be applied. However for categorical responses, intractability of discrete multivariate distributions hampers, at least partly, the development of corresponding likelihood-based methods. Since the publication of the seminal paper of Liang and Zeger (1986),

the generalized estimating equation (GEE) approach has become increasingly important in handling multivariate continuous/discrete responses. There are many attractive points of the GEE. For instance, it is not likelihood-based: only some lower-order moments, such as the mean and variance, of the response need to be specified. Furthermore, one does not even have to model the correlation structure of the response variable correctly; one only needs to use some *working* correlation structure to obtain consistent and asymptotically normal estimates. One convenient choice is the independence model, i.e., the identity matrix serves as the correlation matrix. It has been shown that in many cases the GEE estimates under the independence model (or some other structure) maintain a high efficiency (Zeger (1988), McDonald (1993)).

There are exceptions. For a time-varying covariate, use of the independence model may result in an inefficient estimate (Fitzmaurice (1995)). On the other hand, Pepe and Anderson (1994) have shown that if a non-diagonal working correlation matrix is used, it may lead to seriously biased estimates. Thus either using or not using the independence model may produce bad estimates. It is one of our goals here to demonstrate on a real data set that the straightforward use of a "reasonable" non-diagonal correlation matrix may in fact lead to biased estimates.

It would be desirable if one could choose an appropriate correlation matrix based on available data. In this paper, we propose some new methods to select one from a given set of candidates. The selection criterion is designed to minimize the (estimated) predictive mean squared error (PMSE). For ordinary linear regression, this leads to Akaike's information criterion (AIC) (Akaike (1973)). In our current setting, estimates of PMSE in closed form are not available, and we resort to resampling methods, i.e., to the bootstrap and cross-validation (Efron and Tibshirani (1993), Shao and Tu (1995)).

This paper is organized as follows. In Section 2 we introduce the GEE and the issues involved in using working correlation structures. We propose some new methods for its selection in Section 3. Simulation results are presented in Section 4 to show possible pitfalls of using different working correlation matrices, and the effectiveness of our methods in making an appropriate selection. Section 5 applies the methods to a large clinical trial study. We comment on some practical issues of marginal modeling in Section 6, followed by a brief discussion.

2. Background

We consider the situation in which repeated measurements of the response variable are correlated, as frequently occurs in a longitudinal study. Each subject (or cluster) i $(1 \leq i \leq N)$ is observed at times $t = 1, \ldots, N_i$, with the corresponding response values $Y_i = (Y_{i1}, \ldots, Y_{iN_i})'$ and covariate matrix $X_i =$

$(x_{i1}, \ldots, x_{iN_i})'$. For $i \neq j$, Y_i and Y_j are independent, but generally the components of each Y_i are correlated. The marginal distribution of Y_{it} is specified by a generalized linear model (GLM) (McCullagh and Nelder (1989)): $g(\mu_{it}) = x'_{it}\beta$, where $\mu_{it} = E(Y_{it}|x_{it})$ and g is a given link function. The unknown regression coefficient (vector) β is of primary interest.

The GEE approach estimates β by solving the estimating equations (Liang and Zeger (1986), Prentice (1988)):

$$\sum_{i=1}^{N} D_i' V_i^{-1}(Y_i - \mu_i) = 0, \tag{1}$$

where $D_i = D_i(\beta) = \partial\mu_i(\beta)/\partial\beta'$, and V_i is the *working* covariance matrix of Y_i. V_i can be expressed in terms of a correlation matrix $R(\alpha)$: $V_i = A_i^{1/2} R(\alpha) A_i^{1/2}$, where A_i is a diagonal matrix with elements $var(Y_{it}) = V(\mu_{it})$, specified as functions of the means μ_{it}, α is some unknown parameter. The parameter α can be estimated through moment methods or another set of estimating equations (Prentice (1988)). An attractive point of the GEE approach is that it yields consistent estimator of β, $\hat{\beta} = \hat{\beta}(R)$, even though R is far from a true R_0. Aside from the independence model, $R = I$, other convenient choices include compound symmetry (CS), $R_{ij} = \rho$ for any $i \neq j$, or the first-order autoregressive (AR-1) with $R_{ij} = \rho^{|i-j|}$, where R_{ij} denotes the (i, j)th element of R. The choice of R will influence estimation efficiency: in general, it is more efficient to use an R that is closer to the true correlation. For time-independent covariates, this is not a critical issue. Many studies have shown that $\hat{\beta}$ obtained under the independence model is relatively efficient (Zeger (1988), McDonald (1993)), at least when the correlation between responses is not large. However, for time-varying covariates (i.e., cluster-specific covariates), Fitzmaurice (1993) shows that the resulting estimate from the independence model may be inefficient; its efficiency may be as low as 60% compared to the estimate obtained by using the correct correlation structure. This will be verified in our simulation study in Section 3.

However, this is not the whole story. Pepe and Anderson (1994) point out an implicit assumption behind the GEE approach: the desired statistical properties (e.g., consistency) of $\hat{\beta}$ rely on the unbiasedness of the estimating equations (1). When a non-diagonal working correlation matrix R is used, a sufficient condition for the estimating equations (1) to be unbiased is that

$$E(Y_{it}|x_{it}) = E(Y_{it}|x_{ij}, \;\; j = 1, \ldots, N_i). \tag{2}$$

In practice this assumption may or may not hold. On the other hand, when a diagonal matrix is used, the resulting estimate of β enjoys the aforementioned desirable properties of the GEE. Thus if (2) is violated, one might use the independence model. Note that with time-independent covariates x_{it}, (2) is trivially

satisfied. Although either side of (2) can be modeled, we only call the modeling of $E(Y_{it}|x_{it})$ marginal modeling, unless otherwise specified.

In view of this discussion, it is desirable to choose an appropriate working correlation matrix based on available data. In the next section we propose some resampling methods for doing this.

3. Selecting the Working Correlation Matrix

The problem can be viewed as one of model selection. As in GLM with independent data, it can be done by minimizing the (predictive) mean squared error (Shao and Tu (1995)). Other approaches include an extension of AIC (Pan (2001a)) and minimizing a (predictive) bias of estimating equations (Pan (2001b)).

We denote our current data as $\mathcal{D} = \{(Y_i, X_i), i = 1, \ldots, N\}$, a random sample from some distribution F. If we can repeatedly draw two independent random samples of new observations $\mathcal{D}^{(k)} = \{(Y_i^{(k)}, X_i^{(k)}) : Y_i^{(k)} = (Y_{i1}^{(k)}, \ldots, Y_{iN_i^{(k)}}^{(k)})', X_i^{(k)} = (x_{i1}^{(k)}, \ldots, x_{iN_i^{(k)}}^{(k)})', i = 1, \ldots, N\}$, $k = 1, 2$, from F, it appears reasonable to select a working correlation matrix R that minimizes the predictive mean squared error (PMSE):

$$\begin{aligned} PMSE &= E^{(1)}E^{(2)} \sum_{i=1}^{N} \frac{1}{N} \sum_{t=1}^{N_i^{(2)}} \frac{1}{N_i^{(2)}} \frac{[Y_{it}^{(2)} - \hat{\mu}_{it}^{(2)}(\hat{\beta}(\mathcal{D}^{(1)}, R))]^2}{V(\hat{\mu}_{it}^{(2)}(\hat{\beta}(\mathcal{D}^{(1)}, R)))} \\ &= E^{(1)}E^{(2)} L(\mathcal{D}^{(2)}|\mathcal{D}^{(1)}, R). \end{aligned} \tag{3}$$

Here the expectations $E^{(k)}$ are taken with respect to $\mathcal{D}^{(k)}$, $k = 1, 2$, while $\hat{\mu}_{it}^{(2)}(\hat{\beta}(\mathcal{D}^{(1)}, R)) = g^{-1}(x_{it}^{(2)'}\hat{\beta}(\mathcal{D}^{(1)}, R))$ is the estimated mean of $Y_{it}^{(2)}$ based on $\hat{\beta}(\mathcal{D}^{(1)}, R)$, which is estimated by GEE using the data $\mathcal{D}^{(1)}$ and working correlation matrix R. Implicitly we assume that parameters in R are also estimated in GEE. We use $L(\mathcal{D}^{(2)}|\mathcal{D}^{(1)}, R)$ to denote the squared error obtained by using $\mathcal{D}^{(1)}$ to estimate the regression parameters and using $\mathcal{D}^{(2)}$ to predict the squared error. Note that in (3) we do not consider the correlation within each subject i. Recently Pan and Le (2001) considered an unweighted version of PMSE without the variance term in the denominator of (3). It was shown that in the context of variable selection the unweighted version may have a better performance than the above weighted one.

In practice, we have only one sample of observations $\mathcal{D}$ and cannot calculate PMSE directly. However, resampling methods, such as the bootstrap and cross-validation (Efron and Tibshirani (1993)) can be used to provide effective estimates. A bootstrap sample $\mathcal{D}^*$ is formed by randomly drawing N

observations from $\mathcal{D}$ with replacement. To preserve the correlation structure, the appropriate resampling unit is subject, i.e., (Y_i, X_i) (Rice and Silverman (1991)). Denote a bootstrap sample by $\mathcal{D}^* = \{(Y_i^*, X_i^*) : Y_i^* = (Y_{i1}^*, \ldots, Y_{iN_i^*}^*)', X_i^* = (x_{i1}^*, \ldots, x_{iN_i^*}^*)', i = 1, \ldots, N\}$. A bootstrap estimate (BOOT) of PMSE is

$$BOOT(R) = E^* L(\mathcal{D}|\mathcal{D}^*, R),$$

where the expectation E^* is taken over all bootstrap samples $\mathcal{D}^*$. A similar estimate was proposed by Shao (see Shao and Tu (1995), p.344) for GLM with independent responses.

Another bootstrap estimate (BOOT2), suggested by Efron (1983) in the context of estimating the error rate of a prediction rule, is a sum of the *apparent* error and *excess* error

$$BOOT2(R) = L(\mathcal{D}|\mathcal{D}, R) + E^* \left\{ L(\mathcal{D}|\mathcal{D}^*, R) - L(\mathcal{D}^*|\mathcal{D}^*, R) \right\}.$$

The first term is called apparent error since the same data set is used twice: in both estimating β (and thus μ_i) and predicting the error. The apparent error is a biased estimate of PMSE. The second term corrects the bias of the apparent error in estimating PMSE. An analogous estimate was used in Breiman and Spector (1992) for the usual independent linear regression.

Cross-validation (CV) is another widely used resampling method. CV estimates are in general almost unbiased but may have large variability. Breiman (1995) and Efron and Tibshirani (1997) have suggested using bootstrap to smooth unstable estimators. Here we give a bootstrap-smoothed CV (BCV) estimate of MSE. Let $\mathcal{D}^{*-} = \mathcal{D} - \mathcal{D}^*$ denote the set of observations not appearing in the bootstrap sample $\mathcal{D}^*$. Then the BCV estimate of PMSE is

$$BCV(R) = E^* L(\mathcal{D}^{*-}|\mathcal{D}^*, R).$$

Hence in BCV, as in the usual CV, no observation is used twice: an observation is used either in estimating β (if it is in $\mathcal{D}^*$) or in predicting the error (if in $\mathcal{D}^{*-}$).

In practice, the bootstrap expectation E^* can be approximated by Monte Carlo simulation. If we draw B bootstrap samples $\mathcal{D}^{*b}$, $b = 1, \ldots, B$, then for any function f, $E^* f(\mathcal{D}^*) \approx \sum_{b=1}^{B} f(\mathcal{D}^{*b})/B$ (Efron and Tibshirani (1993), Shao and Tu (1995)). To improve simulation efficiency, the balanced bootstrap can be employed (Efron and Tibshirani (1993)): each subject observation (Y_i, X_i) in $\mathcal{D}$ appears a total of B times in the B bootstrap samples $\mathcal{D}^{*b}$, $b = 1, \ldots, B$. Our experience with the balanced bootstrap (which is used throughout the paper) suggests $B = 25$ works reasonably well.

4. Simulations

4.1. A continuous response variable

Consider the situation in which (2) is violated. Pepe and Anderson (1994) bring in Model 1: $Y_{it} = \alpha Y_{i,t-1} + \beta x_{it} + \epsilon_{it}$, where $Y_{i0} = 0$, x_{it} and ϵ_{it} are independent of each other and of $Y_{i,t-1}$ with a standard normal distribution $N(0,1)$; and Model 2: $Y_{it} = Y_{i,t-1} * \beta x_{it} + \epsilon_{it}$, where $Y_{i0} = 1$ and $\beta = 1$; x_{it}, ϵ_{it} and $Y_{i,t-1}$ are independent of each other, $x_{it} \sim N(1,1)$, and $\epsilon_{it} \sim N(0,1)$. We also consider a random-effects model given as Model 3: $Y_{it} = b_i + \beta x_{it} + \epsilon_{it}$, where b_i, x_{it} and ϵ_{it} are independent of each other and are all distributed as $N(0,1)$.

It can be verified that for Models 1 and 2, the marginal model $E(Y_{it}|x_{it}) = \beta x_{it}$ holds but (2) is not satisfied. Hence we expect that using either CS or AR-1 as the working correlation matrix in GEE will lead to biased estimates of β. Some theoretical results on Model 1 are presented in Emond, Ritz and Oakes (1997) and Pan, Connett and Louis (2000). For Model 3, the above marginal model is also valid, and the true correlation matrix for Y_i is the CS. A simulation study was conducted in `Splus`. We used $N = 50$ or 100, $N_i = 5$, and $\beta = 0.5$ or 1. We considered selecting a working correlation matrix from the independence model, CS and AR-1. The results are shown in Table 1. To facilitate comparison, we also include the results using the unstructured (UN) matrix (i.e., without any restriction on its form) as the working correlation matrix. In addition, the mean squared errors (MSEs) and (some) variances of the regression coefficient estimates are presented to illustrate the bias-variance property. As shown in Section 4.3, the MSE can be decomposed into the sum of variance and bias-square. From Table 1, it appears that the two bootstrap methods work well while BCV is less effective. For Models 1 and 2, both bootstrap methods select the independence model more often than the others, and the resulting estimates have MSEs close to that of the independence model estimates. Also, for Models 1 and 2, it is verified that there is a relatively larger bias contribution to the MSE of the GEE estimate when the working CS or AR-1 structure is used than when the working independence model is used. For Model 3, all three criteria select the correct CS matrix in the majority of cases, though BOOT chose the identity matrix many times. There may be some concerns about the achieved frequency of selecting the best working correlation matrix, since for several set-ups no one method did this correctly more than 60% of trials. However, if we compare the MSEs of the resulting estimates (after selecting the working correlation matrix) with those of the "raw" estimates (without correlation matrix selection), the two bootstrap methods are slightly better than using the working independence model. We suspect that the less than impressive performance of the proposed methods here

may be related to the small MSEs of the GEE estimates under any working correlation structure. For such small MSEs and the given small sample size, the proposed methods may not estimate the PMSE well enough to be able to distinguish the best working correlation structure (CS) from others. For Model 1, when either the sample size is increased to $N = 100$, or the dependence of $E(Y_{it})$ on x_{ij} with $j < t$ is strengthened by increasing α to 1, the performance of all three methods improve. Moreover, when $\alpha = \beta = 1$ (and $n = 50$) in Model 1, the two bootstrap methods chose the independence model in all 100 independent replications, whereas the BCV did so 95 times.

Table 1. Frequency of the working correlation matrix selected by different criteria, and the mean squared error (MSE) and variance (VAR) of the resulting estimate $\hat{\beta}$, from 100 independent replications.

	Model 1				Model 2				Model 3			
	$\alpha = \beta = 0.5, N = 50$				$\beta = 1, N = 50$				$\beta = 0.5, N = 50$			
Criterion	Indp.	CS	AR-1	MSE (VAR)	Indp.	CS	AR-1	MSE (VAR)	Indp.	CS	AR-1	MSE (VAR)
BOOT	60	24	16	.0056	59	13	28	.2976	40	49	11	.0072
*BOOT*2	50	24	26	.0061	50	12	38	.2864	19	53	28	.0061
BCV	29	27	44	.0067	34	14	52	.3259	15	62	23	.0057
Indp.	100	0	0	.0050	100	0	0	.2687	100	0	0	.0079
				(.0045)				(.2450)				(.0070)
CS	0	100	0	.0078	0	100	0	.3421	0	100	0	.0041
				(.0038)				(.1551)				(.0039)
AR-1	0	0	100	.0126	0	0	100	.3197	0	0	100	.0047
				(.0031)				(.1090)				(.0045)
UN	-	-	-	.0142	-	-	-	22.68	-	-	-	.0047
	Model 1				Model 1				Model 1			
	$\alpha = \beta = 0.5, N = 100$				$\alpha = 1, \beta = 0.5, N = 50$				$\alpha = 1, \beta = 1, N = 50$			
Criterion	Indp.	CS	AR-1	MSE (VAR)	Indp.	CS	AR-1	MSE (VAR)	Indp.	CS	AR-1	MSE (VAR)
BOOT	85	13	2	.00265	78	8	14	.0199	100	0	0	.0206
*BOOT*2	71	21	8	.00306	68	11	21	.0241	100	0	0	.0206
BCV	56	27	17	.00349	52	6	42	.0309	95	0	5	.0293
Indp.	100	0	0	.00258	100	0	0	.0146	100	0	0	.0206
				(.00243)				(.0136)				(.0189)
CS	0	100	0	.00512	0	100	0	.0550	0	100	0	.2034
				(.00203)				(.0057)				(.0073)
AR-1	0	0	100	.01065	0	0	100	.0592	0	0	100	.2290
				(.00186)				(.0024)				(.0027)
UN	-	-	-	.01161	-	-	-	.0862	-	-	-	.2285

4.2. A binary response variable

We now consider the model used in Fitzmaurice (1995) to show the inefficiency of the independence model. The response variable is binary, and the

marginal logistic regression model is Model 4: $logit(\mu_{it}) = \beta_0 + \beta_1 x_{it} + \beta_2(t-1)$, $t = 1, 2, 3$, where x_{it} is a dichotomous covariate: $x_{it} = 0$ or 1 with probability 0.5, and $\beta_0 = 0.25 = -\beta_1 = -\beta_2$. We are interested in comparing estimates of β_1. The true correlation matrix is one of the identity, CS or AR-1. We used $\rho = 0.5$ and $N = 50$ or 100. The joint distribution of the Y_i was simulated from Bahadur's (1961) representation (see Fitzmaurice (1995) for more details).

Table 2. Frequency of the working correlation matrix selected by different criteria, and the mean squared error (MSE) and variance (VAR) of the resulting estimate $\hat{\beta}$ in the marginal logistic regression Model 4, from 100 independent replications.

	R_0 =AR-1, $N = 50$				R_0 =CS, $N = 50$				R_0 =Indp., $N = 50$			
Criterion	Indp.	CS	AR-1	MSE (VAR)	Indp.	CS	AR-1	MSE (VAR)	Indp.	CS	AR-1	MSE (VAR)
BOOT	14	30	56	.0774	19	56	25	.0748	48	22	30	.1165
*BOOT*2	12	33	55	.0771	17	52	31	.0722	35	31	34	.1155
BCV	5	32	63	.0780	6	65	29	.0795	40	30	27	.1180
Indp.	100	0	0	.1349	100	0	0	.1346	100	0	0	.1216
				(.1348)				(.1335)				(.1210)
CS	0	100	0	.0895	0	100	0	.0831	0	100	0	.1237
				(.0895)				(.0828)				(.1231)
AR-1	0	0	100	.0805	0	0	100	.0860	0	0	100	.1278
				(.0803)				(.0858)				(.1271)
UN	-	-	-	.0833	-	-	-	.0859	-	-	-	.1317
	R_0 =AR-1, $N = 100$				R_0 =CS, $N = 100$				R_0 =Indp., $N = 100$			
Criterion	Indp.	CS	AR-1	MSE (VAR)	Indp.	CS	AR-1	MSE (VAR)	Indp.	CS	AR-1	MSE (VAR)
BOOT	19	21	60	.0344	21	47	32	.0359	38	30	32	.0479
*BOOT*2	18	23	59	.0342	18	48	34	.0337	30	34	36	.0480
BCV	13	17	70	.0343	12	51	37	.0342	38	25	37	.0480
Indp.	100	0	0	.0549	100	0	0	.0579	100	0	0	.0491
				(.0543)				(.0569)				(.0455)
CS	0	100	0	.0396	0	100	0	.0350	0	100	0	.0502
				(.0394)				(.0346)				(.0468)
AR-1	0	0	100	.0340	0	0	100	.0364	0	0	100	.0506
				(.0338)				(.0361)				(.0472)
UN	-	-	-	.0353	-	-	-	.0365	-	-	-	.0489

The results are shown in Table 2. First, if the true correlation structure R_0 is not the independence model, $\hat{\beta}(I)$ may be inefficient: its relative efficiency against $\hat{\beta}(CS)$ or $\hat{\beta}$(AR-1) may be as low as 60%. However, any one of our proposed methods is most likely to select the correct correlation matrix. The three methods have a close performance: their resulting estimates have almost equal MSEs to those of the estimates obtained using the correct correlation matrix.

Second, when the true correlation structure is the independence model, the GEE estimate from any one of the three working correlation matrices is almost equally efficient. Even though all of our proposed methods select the three working matrices with almost the same frequency, the resulting estimates have MSEs close to those obtained under the correct independence model. This is not surprising since, in this situation, there is essentially no difference in using any one of the three correlation matrices. Note that both the CS and AR-1 models include the independence model ($\rho = 0$).

When the sample size is increased, the frequency of selecting the correct correlation matrix does not increase very much, though the MSE of the resulting estimate is still competitive when compared with using the correct correlation matrix or UN. For Model 4 all the GEE estimates using any working correlation matrix are consistent. Even though $\hat{\beta}(R_0)$ is the most efficient, there is no guarantee that for any given data, $\hat{\beta}(R_0)$ is closer to the true β than other estimate, say $\hat{\beta}(R)$, is. This is analogous to variable selection using AIC. As the sample size increases, AIC will minimize the PMSE but may still choose too large models. Since the motivation of selecting the working correlation matrix is to increase the efficiency of the resulting estimate, this is not of concern as long as the resulting estimate has a small MSE.

Finally we note that, as expected, the MSE contribution from the bias of the regression coefficient estimates for any working correlation matrix is negligible in all cases.

4.3. Performance analysis

We first outline a heuristic argument for using the PMSE in a simplified situation. Suppose that Z_1 and Z_2 are two independent random variables from a distribution with mean μ and variance σ^2. We use E_i to denote the expectation with Z_i, $i = 1, 2$. We also use $\hat{\mu}_1$ to denote an estimate of μ based on Z_1. Then we have

$$\begin{aligned} PMSE &= E_1E_2(Z_2 - \hat{\mu}_1)^2 = E_1E_2(Z_2 - \mu + \mu - \hat{\mu}_1)^2 \\ &= E_1[E_2(Z_2 - \mu)^2 + 2E_2(Z_2 - \mu)(\mu - \hat{\mu}_1) + E_2(\hat{\mu}_1 - \mu)^2] \\ &= \sigma^2 + E_1(\hat{\mu}_1 - \mu)^2. \end{aligned}$$

The last equality is obtained since Z_1 and Z_2 are independent. It is thus obvious that PMSE is minimized when $E_1(\hat{\mu}_1 - \mu)^2$, the MSE of the estimate $\hat{\mu}_1$, is also minimized. In a GLM, we can regard Z_i as a pair of the response variable and covariates (Y_i, X_i), and μ_i depends on the regression coefficient β through $g(\mu_i) = X_i\beta$. By the monotonicity of the link function $g()$ and a first

order Taylor expansion, we can argue that the MSE of $\hat{\mu}_1$ is (approximately) minimized when the MSE of the corresponding regression coefficient estimate $\hat{\beta}_1$ is also minimized.

For correlated data PMSE is defined under the working independence assumption, but this does not imply that it will favor the working independence model in GEE, as demonstrated in our simulations. The reason is that PMSE is based on the *predictive* performance of the estimator in predicting future observations, not the performance of predicting the data used in estimating the regression parameters.

Our simulation results show that (i) BCV is much less effective than the other two bootstrap methods when (2) is violated, but (ii) it performs almost equally well as the other two when (2) holds. The second observation was expected but the first is surprising. In this regard, we recall that the motivation for using CV is to avoid double use of data: any observation is used in estimating either β or PMSE, but not both. Double use of data leads to a downward-biased PMSE estimate. However, it also implies that the evaluation set used in calculating PMSE is smaller, leading to larger variability in CV. Often, including when (2) holds, there is a nice trade-off between the bias and variance in CV. However, when (2) is violated, this is not the case. More specifically, we compare BOOT and BCV. For a given bootstrap sample $\mathcal{D}^*$, BOOT involves evaluating $L(\mathcal{D}|\mathcal{D}^*, R)$ while BCV uses $L(\mathcal{D}^{*-}|\mathcal{D}^*, R)$. Any observation (Y_i^*, X_i^*) in $\mathcal{D}^*$ is used in calculating PMSE through evaluating $L((Y_i^*, X_i^*)|\mathcal{D}^*, R)$ for BOOT, but not for BCV. Suppose Rd is a diagonal correlation matrix and Rnd is a non-diagonal one. Though $L((Y_i^*, X_i^*)|\mathcal{D}^*, R)$ is in general a downward biased PMSE estimate, however, because of the bias of $\hat{\beta}(R_{nd})$ we still expect that more likely $L((Y_i^*, X_i^*)|\mathcal{D}^*, R_{nd}) > L((Y_i^*, X_i^*)|\mathcal{D}^*, R_d)$, leading to a higher chance of obtaining $BOOT(R_{nd}) > BOOT(R_d)$. In contrast, a part of this relevant information is not used in BCV.

4.4. Other simulation results

We tried bootstrap replication number $B = 50$. The results were close to those shown in Tables 1 and 2. We also used the m-out-of-N bootstrap with $m = N/2$ (Shao and Tu (1995)), but its performance was not as good as the usual bootstrap.

5. Example

Taking the Lung Health Study (LHS) (Connett et al. (1993)) data as an example, we first verify that Pepe and Anderson's result is a real issue in practice. Second, we show the effectiveness of our proposed methods in selecting the working correlation matrix.

Specifically, we consider the effect of the change of smoking status on lung function and chronic cough in the LHS. The LHS was a multi-center, randomized controlled clinical trial designed in part to determine whether smoking intervention has a positive effect on the annual rate of decline in lung function. The participants were smokers between the ages 35 and 60 at the beginning of the study. They were randomized into one of three treatment groups: Smoking Intervention plus inhaled ipratropium bromide (SIA), Smoking Intervention and an inhaled placebo (SIP), and Usual Care (UC, no intervention). A behavioral intervention program was provided to all participants in the two intervention groups to encourage and help them quit smoking. The participants were followed for five years. At each annual visit information about changes in smoking habits since the last visit was collected along with other relevant information.

We first take the forced expiratory volume within one second (FEV_1) as the response variable. To ease the comparison between cross-sectional and longitudinal methods, we only include the participants with complete follow-up examinations at five annual visits in our data sets. Three such data sets were formed by taking the first 100, 500 and 1000 participants with five annual examinations. A linear regression model was fitted at each of the five visit years with the following covariates: the current-year smoking-status, treatment group and some baseline characteristics (age, gender, body mass index, body weight, smoking pack-years, cigarettes smoked per day, systolic blood pressure and FEV_1). We are most interested in how the smoking status of a participant at the year of examination influences his/her FEV_1 measure. Previous studies have shown that quitting smoking is associated with a positive effect on FEV_1. From Table 3, it is confirmed that quitting smoking is associated with an increase of FEV_1 by about 0.09 to 0.19 liters. Now we fit a linear model with all the covariates mentioned above plus the visiting year (treated as a categorical variable), using the GEE to combine the results from year 1 to year 5. Three working correlation structures are used: independence model, CS and AR-1 matrices. The last two correlation structures appear plausible considering the likely correlation of the five FEV_1 measures arising from the same participant. It is interesting to note that the point estimates given by GEE/CS or GEE/AR-1 are all *smaller* than any of the cross-sectional estimates. The GEE/Indp. estimates appear to be more reasonable. In this example, it is plausible that condition (2) is violated: given the previous and current years of smoking status, it appears likely that FEV_1 is *not* completely determined by one's current year smoking status. In other words, even though FEV_1 tends to increase if one quits smoking in the current year, it may also depend on *when* the participant stopped smoking.—A five-year sustained quitter is likely to have a different (expected) FEV_1 from a one-year quitter. The independence correlation matrix was selected as the working correlation structure in GEE for each of the three data sets.

Table 3. Estimated regression coefficient (standard error) for the effect of the current year smoking-status on FEV_1 by cross-sectional and longitudinal (GEE) methods. Also included are the GEE estimates of the regression coefficients for the treatment effects, SIA or SIP vs UC group, for $n = 1000$.

	Cross-sectional					Longitudinal		
n	Year 1	Year 2	Year 3	Year 4	Year 5	Indp.	CS	AR-1
100	.1308	.1435	.1266	.2629	.1876	.1641	.0739	.0699
	(.0447)	(.0483)	(.0462)	(.0512)	(.0590)	(.0354)	(.0190)	(.0222)
500	.1145	.1349	.1358	.1812	.1656	.1447	.0781	.0982
	(.0212)	(.0201)	(.0213)	(.0218)	(.0236)	(.0161)	(.0113)	(.0212)
1000	.0951	.1146	.1340	.1693	.1783	.1384	.0815	.0785
	(.0157)	(.0158)	(.0163)	(.0167)	(.0172)	(.0123)	(.0084)	(.0084)
SIA						.0439	.0591	.0570
						(.0161)	(.0160)	(.0158)
SIP						.0060	.0190	.0220
						(.0158)	(.0158)	(.0157)

Note that the biasedness of a regression coefficient estimate may influence other estimates of regression coefficients in the model. In Table 3, we also attach the GEE estimates for the treatment effects for sample size $n = 1000$. It can be seen that there are some differences between the estimates obtained under the working independence model and those under CS/AR-1, though the differences are not dramatic for this data set.

Table 4. Estimated regression coefficient (standard error) for the effect of the current year smoking-status on FEV_1 by the longitudinal (GEE) methods (after considering the interaction between the smoking-status and visiting year and using different working correlation matrix R), with sample size $n = 1000$.

R	Year 1	Year 2	Year 3	Year 4	Year 5
Indp.	.0867	.1123	.1323	.1732	.1805
	(.0150)	(.0154)	(.0158)	(.0159)	(.0164)
CS	.0250	.0514	.0669	.1120	.1361
	(.0122)	(.0109)	(.0110)	(.0111)	(.0116)
AR-1	.0365	.0509	.0602	.1016	.1327
	(.0111)	(.0109)	(.0116)	(.0114)	(.0119)
UN	.0315	.0492	.0564	.0959	.1214
	(.0115)	(.0109)	(.0114)	(.0115)	(.0120)

From Table 3, it appears that the effect of smoking status changes over time. As confirmation we fit a larger marginal model, which treats the visiting year as a categorical variable and includes a two-way interaction term of the visit-

ing year and smoking status, to the data set containing 1000 participants. We can compare the result from cross-sectional analysis (in Table 3) with that from the longitudinal analysis as presented in Table 4. The same phenomenon as observed above persists: the independence model gives results closer to those from the cross-sectional approach, whereas the CS and AR-1 models seem to underestimate the effects of quitting smoking. We also tried an unstructured (UN) working correlation matrix and obtained results similar to the other two non-diagonal working correlation structure methods. Note that because of the presence of the other covariates, the GEE estimates using the working independence model in the longitudinal analysis are not exactly equal to the corresponding ones from the cross-sectional models.

Now we consider the end point as whether the participant reported chronic cough during each year. Two data sets are used with sample sizes $n = 500$ and 1000, respectively. A logistic regression model is fitted with the aforementioned covariates. The results are presented in Table 5. Comparing the estimates from cross-sectional and longitudinal methods, we note that there is a similar trend as shown in Tables 3 and 4. It is possible that both GEE/CS and GEE/AR-1 give downward-biased point estimates of the regression coefficient for the current year smoking-status (i.e., the odds ratio of coughing for a current year smoking quitter vs. a smoker), whereas the GEE estimate under the independence model is more consistent with the cross-sectional estimates. Again our proposed methods all correctly selected the working independence model for each sample size.

Table 5 also shows the GEE estimates for the treatment effects for the sample size $n = 1000$. Again there are some differences between the estimates obtained under the working independence model and those under the CS/AR-1.

Table 5. Estimated regression coefficient (standard error) for the effect of the current year smoking-status on *coughing* by cross-sectional and longitudinal (GEE) methods. Also included are the GEE estimates of the regression coefficients for the treatment effects, SIA or SIP vs UC group, for $n = 1000$.

	Cross-sectional					Longitudinal		
n	Year 1	Year 2	Year 3	Year 4	Year 5	Indp.	CS	AR-1
500	2.351	2.813	1.930	2.615	2.215	2.325	1.776	1.860
	(.398)	(.447)	(.335)	(.388)	(.346)	(.204)	(.186)	(.178)
1000	1.864	2.735	2.209	2.138	2.243	2.209	1.766	1.841
	(.244)	(.312)	(.253)	(.241)	(.240)	(.144)	(.129)	(.126)
SIA						.3100	.3855	.3680
						(.1300)	(.1281)	(.1271)
SIP						.3218	.3787	.3780
						(.1253)	(.1244)	(.1229)

In practice, the data analyst may forget questioning condition (2) and consider using a non-diagonal working correlation matrix due to the strong possibility of existence of correlation of the response variable across the time. This is in fact what motivated our work. Furthermore, even after it is detected that condition (2) is violated, one may still want to find a working correlation structure that works best. Our proposal of using PMSE provides such an approach.

6. More on Marginal Modeling

In this section, we discuss marginal modeling when the time-varying covariates of the same subject are correlated. In Pepe and Anderson's orginal paper, only independent time-varying covariates were discussed.

Consider the following model where (2) is violated:

$$E(Y_{it}|x_{ij}, j = 1, \ldots, N_i) = \alpha + \sum_{j=1}^{t} \beta_j x_{ij}, \qquad t = 1, \ldots, N_i.$$

So far we have assumed that all x_{ij}'s are independent. In practice, it is more likely that $x_{i1}, \ldots, x_{iN_i}$ are correlated. It is then natural to ask what is the corresponding marginal model. In general, denote $E(x_{ij}|x_{ik}) = f_{jk}(x_{ik})$ for some function f_{jk} for any given $j \neq k$. Then the marginal model is $E(Y_{it}|x_{it}) = \alpha + \sum_{j=1}^{t-1} \beta_j f_{jt}(x_{it}) + \beta_t x_{it}$, which may not be a linear model in the first place. However, if for any $j \neq k$,

$$E(x_{ij}|x_{ik}) = \alpha_{jk} + \gamma_{jk} x_{ik}, \tag{4}$$

then the marginal model has the usual form

$$E(Y_{it}|x_{it}) = \alpha_t + \beta_t^* x_{it}, \tag{5}$$

where $\alpha_t = \alpha + \sum_{j=1}^{t-1} \beta_j \alpha_{jt}$ and $\beta_t^* = \sum_{j=1}^{t-1} \beta_j \gamma_{jt} + \beta_t$. Hence it is required to fit the marginal model using time-varying intercepts and slopes.

We believe that (4) is a convenient but still reasonable assumption. When $(x_{i1}, x_{i2}, \ldots, x_{iN_i})$ has a multivariate normal distribution, (4) is satisfied. Thus transforming continuous covariates to make them appear normal will be helpful in marginal modeling. Many multivariate discrete distributions (Johnson and Kotz (1969), Chapter 11) have properties satisfying (2). In our example, x_{ij} corresponds to the smoking status of person i at year j. Since each x_{ij} is binary, it appears reasonable to model it as a binomial (or Bernoulli) $Bin(1, p_i)$. To introduce correlation among $(x_{i1}, \ldots, x_{i5})$, we further model p_i as a random variable from a beta distribution. In other words, if x_{it}'s are modeled as from a beta-binomial distribution, it can be shown following Johnson and Kotz ((1969),

p.309) that (4) is satisfied with constant $\alpha_{jk} = \alpha_0$ and $\gamma_{jk} = \gamma_0$ for any $j \neq k$. Both α_0 and γ_0 depend on the parameters of the specified beta-binomial distribution.

7. Discussion

In this article we have discussed selecting the working correlation structure in GEE for GLMs with dependent response data. We motivated our study mainly in light of Pepe and Anderson's result: using a non-diagonal working correlation matrix may lead to biased estimates of regression parameters if the implicit assumption (2) for GEE is violated.

A seemingly straightforward solution to the issue is to model $E(Y_{it}|x_{ij}, j = 1, \ldots, N_i)$, rather than $E(Y_{it}|x_{it})$. Pepe and Anderson argued that in some situations the latter is preferred. Pepe, Whitaker and Seidel (1999) demonstrated one of its interesting applications. In principle, with the presence of time-varying covariates, condition (2) should be checked. If it is clear that (2) is violated, then the working independence model should be used. However, one may not always check (2). Our proposal to select an appropriate working correlation matrix provides an alternative. Generally, selecting and using an appropriate working correlation structure may improve estimation efficiency.

Acknowledgement

The authors wish to thank Tom Louis and Jim Neaton for helpful discussions and comments. They also thank a referee, an associate editor and the editor for careful reading and constructive comments. The authors were supported by NIH grants.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. 2*nd International Symposium on Information Theory* (Edited by B. N. Petrov and F. Csaki), 267-281. Akademiai Kiado, Budapest.

Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. In *Studies in Item Analysis and Prediction* (Edited by H. Solomon), 158-168. Stanford Mathematical Studies in the Social Sciences VI. Stanford University Press, Stanford.

Breiman, L. (1996). Bagging predictors. *Machine Learning* **26**, 123-140.

Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression: the X random case. *Internat. Statist. Rev.* **60**, 291-319.

Connett, J. E., Kusek, J. W., Bailey, W. C., O'Hara, P. and Wu, M. for the Lung Health Study Research Group (1993). Design of the Lung Health Study: a randomized clinical trial of early intervention for chronic obstructive pulmonary disease. *Controlled Clinical Trials* **14**, 3S-19S.

Efron, B. (1983). Estimating the error rate of a prediction rule: some improvements on cross-validation. *J. Amer. Statist. Assoc.* **78**, 316-331.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, London.

Efron, B. and Tibshirani, R. J. (1997). Improvements on cross-validation: the 0.632+ bootstrap method. *J. Amer. Statist. Assoc.* **92**, 548-560.

Emond, M. J., Ritz, J. and Oakes, D. (1997). Bias in GEE estimates from misspecified models for longitudinal data. *Comm. Statist. Ser. A* **26**, 15-32.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multiple multivariate binary data. *Biometrics* **51**, 309-317.

Johnson, N. I. and Kotz, S. (1969). *Discrete Distributions.* Houghton Mifflin, Boston.

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

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* 2nd edition. Chapman and Hall, London.

McDonald, B. W. (1993). Estimating logistic regression parameters for bivariate binary data. *J. Roy. Statist. Soc. Ser. B* **55**, 391-397.

Pan, W. (2001a). Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120-125.

Pan, W. (2001b). Model selection in estimating equations. *Biometrics* **57**, 529-534.

Pan, W., Connett, J. E. and Louis, T. A. (2000). A note on marginal linear regression with correlated response data. *Amer. Statist.* **54**, 191-195

Pan, W. and Le, C. T. (2001). Bootstrap model selection in generalized linear models. *J. Agricultural, Biological and Environmental Statist.* **6**, 49-61.

Pepe, M. S. and Anderson, G. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Comm. Statist. Ser. B* **23**, 939-951.

Pepe, M. S., Whitaker, R. C. and Seidel, K. (1999). Estimating and comparing univariate associations with application to the prediction of adult obesity. *Statist. in Medicine* **18**, 163-173.

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

Rice, J. R. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53**, 233-243.

Shao, J. and Tu, D. S. (1995). *The Jackknife and Bootstrap.* Springer, New York.

Zeger, S. L. (1988). The analysis of discrete longitudinal data: Commentary. *Statist. in Medicine* **7**, 161-168.

Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, MMC 303, 420 Delaware Street SE, Minneapolis, MN 55455-0378, U.S.A.

E-mail: weip@biostat.umn.edu

E-mail: john-c@biostat.umn.edu

(Received October 2000; accepted October 2001)