On the robust variance estimator in generalised estimating equations

By WEI PAN

Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

Summary

The variance matrix of the estimated regression coefficient in Liang and Zeger’s generalised estimating equation approach can be consistently estimated by the so-called sandwich or robust estimator. In this note, we propose a modification to Liang and Zeger’s prescription for implementing the robust variance estimator. Analytical and numerical evidence shows the superior performance of our proposal.

Some key words: Generalised estimating equation; Generalised linear model; Longitudinal data; Sandwich estimator.

1. Introduction

In their seminal papers (Liang & Zeger, 1986; Zeger & Liang 1986), Liang and Zeger extended generalised linear models (McCullagh & Nelder, 1989) to handle longitudinal and other dependent response data. Their work can be regarded as a multivariate extension of the quasilikelihood method of Wedderburn (1974). In longitudinal studies, each subject $i$ may have several measurements of a response $y_{it}$ and a covariate vector $x_{it}$, for $t = 1, ..., n_i$ and $i = 1, ..., K$. Here for simplicity we consider the situation that all $n_i = n$ as in Liang & Zeger (1986); for varying $n_i$, it can be handled in a similar way. Denote $(y_{i1}, ..., y_{in})'$ by $Y_i$ and $(x'_{i1}, ..., x'_{in})'$ by $X_i$. The marginal model specifies a relationship between the marginal mean $E(Y_i|X_i) = \mu_i$
and the covariate $X_i$ through a generalised linear model: $g(\mu_i) = X_i \beta$, where $\beta$ is an unknown $p$-vector of regression coefficients, and $g$ is a known link function; the marginal variance is $\text{var}(y_{it} | x_{it}) = v(\mu_{it}) \phi$, where $v$ is a known function and $\phi$ is a scale parameter which may need to be estimated; and the within-subject correlation matrix $\text{corr}(Y_i)$ is $R_0(\alpha)$, where the structure of $R_0$ is in general unknown and may depend on a parameter $\alpha$. It is assumed throughout that $Y_i$ and $Y_j$ are independent for any $i \neq j$. An attractive point of Liang and Zeger’s generalised estimating equation approach is that one does not need to specify $R_0$ correctly; instead, one can use some working correlation matrix $R_W(\alpha)$; see Crowder (1995) for other related issues. The approach is non-likelihood-based: the asymptotic validity of the estimators depends only on the correct specification of the mean of $y_{it}$. Let $A_i = \text{diag}\{v(\mu_{i1}), ..., v(\mu_{in})\}$ and let the working covariance matrix be $V_i = \phi A_i^{1/2} R_W(\alpha) A_i^{1/2}$. Then the generalised estimating equation approach estimates $\beta$ by solving the following estimating equations:

$$
\sum_{i=1}^{K} D_i' V_i^{-1} S_i = 0, 
$$

where $D_i = \partial \mu_i / \partial \beta'$ and $S_i = Y_i - \mu_i$. Provided that we have $K^{1/2}$-consistent estimators $\hat{\alpha}$ and $\hat{\phi}$, under mild regularity conditions, Liang and Zeger show that $\hat{\beta}$, the solution of (1), is consistent and asymptotically normal; the covariance matrix of $\hat{\beta}$ can be consistently estimated by the so-called sandwich or robust estimator

$$
V_G = \left( \sum_{i=1}^{K} D_i' V_i^{-1} D_i \right)^{-1} \left( \sum_{i=1}^{K} D_i' V_i^{-1} \text{cov}(Y_i) V_i^{-1} D_i \right) \left( \sum_{i=1}^{K} D_i' V_i^{-1} D_i \right)^{-1},
$$

where $\beta$, $\alpha$ and $\phi$ are replaced by their estimators; Liang and Zeger propose to estimate $\text{cov}(Y_i)$ by $S_i S_i'$. We denote the $V_G$ thus obtained by $V_{LZ}$. Usually, the middle factor of $V_{LZ}$ divided by $K$, $\sum_{i=1}^{K} D_i' V_i^{-1} S_i S_i' V_i^{-1} D_i / K$, is regarded as an empirical
variance estimator of the estimating function in (1). Here we argue that $S_i S'_i$ is not an optimal estimator of $\text{cov}(Y_i)$: it is neither consistent nor efficient since it is based on the data from only one subject, and we propose a modification to it next.

2. Modification To The Sandwich Estimator

Liang and Zeger also suggest that $R_0$ can be estimated by

$$R_U = \frac{1}{\phi K} \sum_{i=1}^{K} A_i^{-1/2} S_i S'_i A_i^{-1/2},$$

(3)

where the unknown $\beta$ and $\phi$ are replaced by their estimators $\hat{\beta}$ and $\hat{\phi}$. Note that $R_U$ is obtained without any parametric specification of the structure of the correlation matrix $R_0$, and hence is also called an unstructured estimator. We propose that $\text{cov}(Y_i)$ be estimated by

$$W_i = \phi A_i^{1/2} R_U A_i^{1/2} = A_i^{1/2} \left( \sum_{i=1}^{K} A_i^{-1/2} S_i S'_i A_i^{-1/2} / K \right) A_i^{1/2}.$$

(4)

Replacing $\text{cov}(Y_i)$ in (2) by $W_i$, we obtain a new covariance estimator $V_N$. Note that $W_i$ does not depend on $\phi$.

The following two additional assumptions are needed to guarantee the asymptotic validity of the new estimator $V_N$.

Assumption 1: The marginal variance of $y_{ii}$ needs to be modelled correctly.

Assumption 2: There is a common correlation structure $R_0$ across all subjects.

We believe that these two assumptions are often reasonable. The modelling of marginal variance in the current context is the same as that for independent data, which has been extensively discussed in the literature. The use of analogues of Assumption 2 is also popular in practice, for instance in linear mixed-effects models for continuous correlated response data. Our view is that, as with any modelling...
assumption, Assumption 1 or Assumption 2 may correspond to a good approximation unless there is strong evidence against it. This is particularly relevant when the sample size $K$ is not large; see Agresti (1990, §6.4.4) for a nice discussion on the advantage of modelling. Otherwise, it is of less concern to use Liang and Zeger’s $V_{LZ}$.

Even if it is suspected that Assumption 2 does not hold, we may classify subjects into several groups such that all $Y_i$ in the same group have the same correlation matrix, and each $R_U$ can be accordingly obtained. In the extreme situation where each $Y_i$ has a different correlation structure, $R_{U_i} = A_i^{-1/2}S_iS_i' A_i^{-1/2}/\phi$ and our estimator $W_i$ reduces to Liang and Zeger’s $S_iS_i'$. Unless otherwise specified, $V_N$ is obtained under the assumption that there is a common correlation structure across all subjects.

Since our estimator $W_i$ is obtained by pooling observations across different subjects, whereas Liang and Zeger’s $S_iS_i'$ is based on the observation from subject $i$ only, we expect our estimator to be more efficient than Liang and Zeger’s. Here we provide a partial justification by comparing their variabilities, and we treat the $D_i, V_i$ and $A_i$ as fixed. Since both estimators share the same two outside factors in (2), we need only compare their middle factors

$$M_N = \sum_{i=1}^{K} D_i' V_i^{-1} W_i V_i^{-1} D_i \quad \text{and} \quad M_{LZ} = \sum_{i=1}^{K} D_i' V_i^{-1} S_i S_i' V_i^{-1} D_i.$$  

For any matrix $B$, define the operator $\text{vec}(B)$ as that of stacking the columns of $B$ together to obtain a vector. In the Appendix we prove the following theorem.

**Theorem 1.** Under mild regularity conditions, $\text{cov}\{\text{vec}(M_{LZ})\} - \text{cov}\{\text{vec}(M_N)\}$ is nonnegative definite with probability tending to 1 as $K \to \infty$.

Therefore, asymptotically, we have that $\text{cov}\{\text{vec}(V_{LZ})\} \geq \text{cov}\{\text{vec}(V_N)\}$. Although we can only prove the above result asymptotically, the simulations below show that the result seems to hold also for small $K$. 
The efficiency gain of the new estimator is achieved at the cost of requiring the two additional assumptions. If they do not hold, asymptotically correct inference about $\beta$ can still be made using $V_N$ according to the following result.

**Theorem 2.** Under mild regularity conditions and provided that the marginal mean is correctly specified, the Wald statistic $(\hat{\beta} - \beta)^t V_N^{-1} (\hat{\beta} - \beta)$ asymptotically has a distribution $\sum_{j=1}^{p} c_j \chi_j^2$, where $\chi_j^2$'s are $p$ independent chi-squared random variables with 1 degree of freedom, and the $c_j$'s can be consistently estimated by the eigenvalues of $V_N^{-1}V_{LZ}$.

The proof of Theorem 2 is given in the Appendix. As in Rotnitzky & Jewell (1990), the $c_j$'s provide an asymptotic check on the modelling Assumptions 1 and 2. If the assumptions are correct, both $V_N$ and $V_{LZ}$ are consistent and thus all $c_j$'s should be 1's, leading to the usual chi-squared distribution of the Wald statistic. The mean and variance of the Wald statistic are $s_1 = \sum_{j=1}^{p} c_j$ and $s_2 = 2 \sum_{j=1}^{p} c_j^2$, which can be estimated as $\text{tr}(V_N^{-1}V_{LZ})$ and $2\text{tr}\{(V_N^{-1}V_{LZ})^2\}$ respectively. Following the suggestion of Rotnitzky and Jewell, one can compare $(s_1, s_2)$ with $(p, 2p)$ to gain some idea about the validity of the two assumptions. However, as Rotnitzky and Jewell pointed out, a probability statement about $s_1$ and $s_2$ is complicated, because of the difficulty in determining the null distribution of $V_N^{-1}V_{LZ}$.

Generally the diagonal elements of $R_U$ need not be 1’s. Since $R_U$ is estimating a correlation matrix, we may just use nondiagonal elements of $R_U$ and stipulate all diagonal elements of $R_U$ as 1’s. Then, even in the extreme situation of estimating each $\text{corr}(Y_i)$ separately, our $W_i$ is different from the $S_i S_i'$ of Liang and Zeger. The diagonal elements of $W_i$ are equal to $\sigma^2 \phi$, which are natural plug-in estimators of $\text{var}(y_{ii}) = \sigma^2 \phi$. In contrast, in the version of Liang and Zeger’s they are $(y_{ii} - \hat{\mu}_{ii})^2$, each of which is an empirical variance estimator based on only one observation.
3. Numerical examples

In this section, we compare the performance of the two robust variance estimators \( V_{LZ} \) and \( V_N \) through simulation. First, we consider a linear random-effects model:

\[
y_{ij} = \beta_0 + x_{ij} \beta_1 + b_i + e_{ij},
\]

where the \( x_{ij}, b_i \) and \( e_{ij} \) are independently and identically distributed as \( N(0, 1) \), and they are independent of each other, for \( j = 1, 2, 3 \) and \( i = 1, \ldots, K \). We fix \( \beta_0 = 1 \) and set \( \beta_1 = 0 \) or \( 1 \), and \( K = 10 \) or \( 40 \). Note that for this model, Assumptions 1 and 2 hold.

Two working correlation matrices are used: one is the identity matrix \( R_W = I \); the other is the exchangeable or compound symmetry matrix with diagonal elements \( R_W(i, i) = 1 \) and nondiagonal elements \( R_W(i, j) = \rho \) for \( i \neq j \). Our simulation was conducted in S-Plus. In particular, we used the S-Plus function `gee()` to obtain the estimated regression coefficient \( \hat{\beta}_1 \) and its robust variance estimate \( V_{LZ} \). The results in Table 1 are based on 1000 independent simulations for each set-up. It is seen that \( V_N \) is closer than is \( V_{LZ} \) to the true variance \( V(\hat{\beta}_1) \), as estimated by the sample variance of \( \hat{\beta}_1 \) from 1000 simulations. In agreement with Theorem 1, we also see that the variance of \( V_N \) is smaller than that of \( V_{LZ} \). Not surprisingly, using the two variance estimators will have different implications for statistical inference. Here we consider the size of the two-sided \( z \)-test for \( H_0 : \beta_1 = 0 \) against \( H_1 : \beta \neq 0 \). The \( z \)-statistic is based on either of the two variance estimators: \( z = \hat{\beta}_1 / (V_{LZ})^{1/2} \) or \( z = \hat{\beta}_1 / (V_N)^{1/2} \). It is clear that the \( z \)-test based on our new variance estimator has size closer to the nominal level than does that based on \( V_{LZ} \).

The second model we consider is a random-effects logistic model:

\[
\text{logit}(\mu_{ij} | b_i) = \beta_0 + x_{ij} \beta_1 + b_i,
\]
where $x_{ij}$ and $b_i$ are independently and identically distributed as $N(0,1)$, and they are independent of each other; conditional on $b_i$, $y_{ij} \sim Bi(1, \mu_{ij})$, for $j = 1, 2, 3$ and $i = 1, \ldots, K$. For nonlinear models, random-effects models may not be equivalent to any marginal model, but the above logistic-normal random-effects model can be well approximated by a corresponding marginal logistic model (Zeger et al., 1988). We fix $\beta_0 = 0$ and set $\beta_1 = 0$ or $1$, and $K = 20$ or $40$.

As a result of the well-known fact that the correlation of binary responses is restricted by their marginal means (Prentice, 1988; Diggle et al., 1994, p. 133), it is likely that there is no common correlation matrix $R_0$ across all subjects for binary logistic model. We want to use this scenario to demonstrate that, even if Assumptions 1 and 2 are slightly violated, our proposal may still have better performance. The results based on 1000 simulations are presented in Table 2. The same conclusion can be drawn: our variance estimator of $\hat{\beta}_1$ performs better than that of Liang and Zeger. Note also that the variance of $V_N$ is smaller than that of $V_{LZ}$.

In all of the above simulations, we assumed that all $\text{corr}(Y_i)$ are the same. Now we allow the $\text{corr}(Y_i)$ to differ. Under the working independence model, i.e. $R_W = I$, Liang and Zeger’s estimator is unchanged, and our estimator, obtained by setting the diagonal elements of $R_W$ equal to 1, is presented in Table 3. It appears that our estimator still has a slight edge.

**Acknowledgement**

The author would like to thank John Connett, Chap Le and Tom Louis for interesting discussions. The author is grateful to a referee and the editor for detailed and helpful comments. This research was partially supported by a University of Minnesota Grant-in-aid.
Appendix

Sketch proofs

Proof of Theorem 1. We need to use the Kronecker product $\otimes$ and vec() operations (Vonesh & Chinchilli, 1997, p.12). Since
\[
\text{vec}(M_{LZ}) = \sum_{i=1}^{K} \{(D_i^TV_i^{-1}) \otimes (D_i^TV_i^{-1})\} \text{vec}(S_iS_i'),
\]
we have
\[
\text{cov}\{\text{vec}(M_{LZ})\} = \sum_{i=1}^{K} C_i\Omega_iC_i',
\]
where $C_i = (D_i^TV_i^{-1}) \otimes (D_i^TV_i^{-1})$ and $\Omega_i = \text{cov}\{\text{vec}(S_iS_i')\}$. Similarly,
\[
\text{vec}(M_N) = \sum_{i=1}^{K} C_i\text{vec}\{A_i^{1/2}\sum_{j=1}^{K} A_j^{-1/2}S_jS_j'A_j^{-1/2}A_i^{1/2}\}
\]
and hence
\[
\text{cov}\{\text{vec}(M_N)\} = \sum_{i=1}^{K} C_i\{F_i \sum_{j=1}^{K} \left(\frac{1}{K^2}F_j^{-1}\Omega_jF_j^{-1}\right)F_i\}C_i',
\]
where $F_i = A_i^{1/2} \otimes A_i^{1/2}$. Thus
\[
\text{cov}\{\text{vec}(M_{LZ})\} - \text{cov}\{\text{vec}(M_N)\} = \sum_{i=1}^{K} C_i \left(\Omega_i - F_i \frac{\sum_{j=1}^{K} F_j^{-1}\Omega_jF_j^{-1}}{K^2}F_i\right)C_i'.
\]
Under suitable conditions, $\sum_{j=1}^{K} F_j^{-1}\Omega_jF_j^{-1}/K$ tends to some matrix, $Q$, say, with probability 1 as $K \to \infty$. Thus the second term in the bracket will tend to 0 with probability 1. Note also that in general the nonnegative definite covariance matrix $\Omega_i$ is not a zero matrix. The theorem follows immediately.

Proof of Theorem 2. According to Theorem 2 of Liang & Zeger (1986), we know that $K^{1/2}(\hat{\beta} - \beta)$ asymptotically has a normal distribution with mean 0 and covariance matrix $V_G$, which can be consistently estimated by $V_{LZ}$. Using the well-known property of quadratic forms of normal random variables (Johnson & Kotz, 1970, p. 150), we prove the theorem.
References


Table 1: Means with estimated standard deviations in parentheses of the variance estimates of $\hat{\beta}_1$, and sizes of the 0.05-level, and in parentheses the 0.10-level, two-sided z-test, in the linear mixed-effects model. Results are based on 1000 independent replications.

<table>
<thead>
<tr>
<th>$R_W$</th>
<th>Set-up</th>
<th>$\beta_1$</th>
<th>$K$</th>
<th>$V(\hat{\beta}_1)$</th>
<th>$V_{LZ}$</th>
<th>$V_N$</th>
<th>$V_{LZ}$</th>
<th>$V_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I$</td>
<td></td>
<td>0</td>
<td>10</td>
<td>0.0719</td>
<td>0.0570 (0.0389)</td>
<td>0.0634 (0.0307)</td>
<td>0.128 (0.185)</td>
<td>0.078 (0.131)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>40</td>
<td>0.0168</td>
<td>0.0160 (0.0056)</td>
<td>0.0166 (0.0037)</td>
<td>0.060 (0.116)</td>
<td>0.055 (0.105)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>10</td>
<td>0.0719</td>
<td>0.0570 (0.0389)</td>
<td>0.0634 (0.0307)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>40</td>
<td>0.0168</td>
<td>0.0160 (0.0056)</td>
<td>0.0166 (0.0037)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CS$</td>
<td></td>
<td>0</td>
<td>10</td>
<td>0.0533</td>
<td>0.0389 (0.0251)</td>
<td>0.0434 (0.0189)</td>
<td>0.130 (0.200)</td>
<td>0.088 (0.152)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>40</td>
<td>0.0116</td>
<td>0.0108 (0.0035)</td>
<td>0.0111 (0.0022)</td>
<td>0.068 (0.114)</td>
<td>0.059 (0.109)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>10</td>
<td>0.0533</td>
<td>0.0389 (0.0251)</td>
<td>0.0434 (0.0189)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>40</td>
<td>0.0116</td>
<td>0.0108 (0.0035)</td>
<td>0.0111 (0.0022)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CS: Compound symmetry
LZ: Liang and Zeger’s method
N: New method
e.s.d.: Estimated standard deviation
Table 2: Means with estimated standard deviations in parentheses of the variance estimates of $\hat{\beta}_1$, and sizes of the 0.05-level, and in parentheses the 0.10-level, two-sided $z$-test, in the logistic mixed-effects model. Results are based on 1000 independent replications.

<table>
<thead>
<tr>
<th>$R_W$</th>
<th>$\beta_1$</th>
<th>$K$</th>
<th>$V(\hat{\beta}_1)$</th>
<th>$V_{LZ}$</th>
<th>$V_N$</th>
<th>$V_{LZ}$</th>
<th>$V_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0</td>
<td>20</td>
<td>0.0814</td>
<td>0.0710 (0.0240)</td>
<td>0.0744 (0.0160)</td>
<td>0.068 (0.124)</td>
<td>0.049 (0.107)</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>40</td>
<td>0.0375</td>
<td>0.0343 (0.0081)</td>
<td>0.0352 (0.0050)</td>
<td>0.068 (0.114)</td>
<td>0.058 (0.102)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>20</td>
<td>0.1346</td>
<td>0.1105 (0.0584)</td>
<td>0.1173 (0.0487)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>40</td>
<td>0.0569</td>
<td>0.0526 (0.0164)</td>
<td>0.0541 (0.0120)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td>0</td>
<td>20</td>
<td>0.0788</td>
<td>0.0655 (0.0245)</td>
<td>0.0690 (0.0163)</td>
<td>0.068 (0.135)</td>
<td>0.048 (0.112)</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>40</td>
<td>0.0356</td>
<td>0.0322 (0.0083)</td>
<td>0.0331 (0.0052)</td>
<td>0.066 (0.107)</td>
<td>0.052 (0.098)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>20</td>
<td>0.1315</td>
<td>0.1056 (0.0573)</td>
<td>0.1098 (0.0463)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>40</td>
<td>0.0555</td>
<td>0.0505 (0.0161)</td>
<td>0.0515 (0.0120)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CS: Compound symmetry
LZ: Liang and Zeger’s method
N: New method
e.s.d.: Estimated standard deviation
Table 3: Means with estimated standard deviations in parentheses of the variance estimates of \( \hat{\beta}_1 \), and sizes of the 0.05-level, and in parentheses the 0.10-level, two-sided z-test, in the logistic mixed-effects model. For the proposed new variance estimator, it is assumed that each subject has its own correlation matrix. Results are based on 1000 independent replications.

<table>
<thead>
<tr>
<th>( R_W )</th>
<th>Set-up</th>
<th>Means and e.s.d.’s</th>
<th>Sizes of z-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( V(\hat{\beta}_1) )</td>
<td>( V_{LZ} )</td>
</tr>
<tr>
<td>1</td>
<td>0 20</td>
<td>0.0814</td>
<td>0.0710 (0.0240)</td>
</tr>
<tr>
<td></td>
<td>0 40</td>
<td>0.0375</td>
<td>0.0343 (0.0081)</td>
</tr>
<tr>
<td></td>
<td>1 20</td>
<td>0.1346</td>
<td>0.1105 (0.0584)</td>
</tr>
<tr>
<td></td>
<td>1 40</td>
<td>0.0569</td>
<td>0.0526 (0.0164)</td>
</tr>
</tbody>
</table>

CS: Compound symmetry
LZ: Liang and Zeger’s method
N: New method
e.s.d.: Estimated standard deviation