Biometrika (2001), **88,** 3, *pp.* 901–906

Printed in Great Britain

On the robust variance estimator in generalised estimating equations

BY WEI PAN

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

weip@biostat.umn.edu

SUMMARY

The variance matrix of the estimated regression coefficient in the Liang–Zeger generalised estimating equation approach can be consistently estimated by the so-called sandwich or robust estimator. In this note, we propose a modification to the Liang–Zeger 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 $t = 1, \ldots, n_i$ and $i = 1, \ldots, K$. Here for simplicity we consider the situation that all $n_i = n$ as in Liang & Zeger (1986); the case of varying n_i can be handled in a similar way. Denote $(y_{i1}, \ldots, y_{in})'$ by Y_i and $(x_{i1}', \ldots, x_{in}')'$ 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 β is an unknown p-vector of regression coefficients, and g is a known link function; the marginal variance is $\mathrm{var}(y_{it} | x_{it}) = v(\mu_{it})\phi$, where v is a known function and ϕ is a scale parameter which may need to be estimated; and the within-subject correlation matrix $\mathrm{corr}(Y_i)$ is $R_0(\alpha)$, where the structure of R_0 is in general unknown and may depend on a parameter α. It is assumed throughout that Y_i and Y_j are independent for any $i \neq j$. An attractive point of the generalised estimating equation approach is that one does not need to specify R_0 correctly. Instead, one can use some working correlation matrix $R_{\mathrm{W}}(\alpha)$; 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 $A_i = \mathrm{diag}\{v(\mu_{i1}), \ldots, v(\mu_{in})\}$ and let the working covariance matrix be $V_i = \phi A_i^{1/2} R_{\mathrm{W}}(\alpha) A_i^{1/2}$. Then the generalised estimating equation approach estimates β by solving the following estimating equations:

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

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

$$V_G = \left(\sum_{i=1}^{K} D_i' V_i^{-1} D_i \right)^{-1} \left\{ \sum_{i=1}^{K} D_i' V_i^{-1} \operatorname{cov}(Y_i) V_i^{-1} D_i \right\} \left(\sum_{i=1}^{K} D_i' V_i^{-1} D_i \right)^{-1}, \quad (2)$$

where β, α and ϕ are replaced by their estimators; Liang and Zeger propose to estimate $\operatorname{cov}(Y_i)$ 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 $\operatorname{cov}(Y_i)$: 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 suggest that R_0 can be estimated by

$$R_U = \frac{1}{\phi K} \sum_{i=1}^{K} A_i^{-1/2} S_i S_i' A_i^{-1/2}, \quad (3)$$

where the unknown β and ϕ 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 $\operatorname{cov}(Y_i)$ be estimated by

$$W_i = \phi A_i^{1/2} R_U A_i^{1/2} = A_i^{1/2} \left(\sum_{i=1}^{K} A_i^{-1/2} S_i S_i' A_i^{-1/2} / K \right) A_i^{1/2}. \quad (4)$$

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

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_{it} 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 the Liang–Zeger 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, $R_{Ui} = A_i^{-1/2} S_i S_i' A_i^{-1/2} / \phi$ and our estimator W_i reduces to the Liang–Zeger $S_i S_i'$. 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 the Liang–Zeger $S_i S_i'$ is based on the observation from subject i only, we expect our estimator to be more efficient than the Liang–Zeger. 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 compare only their middle factors,

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

For any matrix B, define the operator $\operatorname{vec}(B)$ 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,* $\mathrm{cov}\{\mathrm{vec}(M_{\mathrm{LZ}})\} - \mathrm{cov}\{\mathrm{vec}(M_{\mathrm{N}})\}$ *is nonnegative definite with probability tending to* 1 *as* $K \to \infty$.

Therefore, asymptotically, we have that $\mathrm{cov}\{\mathrm{vec}(V_{\mathrm{LZ}})\} \geqslant \mathrm{cov}\{\mathrm{vec}(V_{\mathrm{N}})\}$. 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 β 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* $(\hat{\beta} - \beta)' V_{\mathrm{N}}^{-1} (\hat{\beta} - \beta)$ *asymptotically has a distribution* $\sum_{j=1}^{p} c_j \chi_j^2$, *where the* χ_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* $V_{\mathrm{N}}^{-1} V_{\mathrm{LZ}}$.

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 $\mathrm{tr}(V_{\mathrm{N}}^{-1} V_{\mathrm{LZ}})$ and $2\,\mathrm{tr}\{(V_{\mathrm{N}}^{-1} V_{\mathrm{LZ}})^2\}$ respectively. Following the suggestion of Rotnitzky & 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 & Jewell pointed out, a probability statement about s_1 and s_2 is complicated, because of the difficulty in determining the null distribution of $V_{\mathrm{N}}^{-1} V_{\mathrm{LZ}}$.

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 $\mathrm{corr}(Y_i)$ separately, our W_i is different from the Liang–Zeger $S_i S_i'$. The diagonal elements of W_i are equal to $v(\hat{\mu}_{it})\hat{\phi}$, which are natural plug-in estimators of $\mathrm{var}(y_{it}) = v(\mu_{it})\phi$. In contrast, in the Liang–Zeger version they are $(y_{it} - \hat{\mu}_{it})^2$, 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 $R_{\mathrm{W}} = I$; the other is the exchangeable or compound symmetry matrix with diagonal elements $R_{\mathrm{W}}(i, i) = 1$ and nondiagonal elements $R_{\mathrm{W}}(i, j) = \rho$ 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: $z = \hat{\beta}_1/(V_{\mathrm{LZ}})^{1/2}$ or $z = \hat{\beta}_1/(V_{\mathrm{N}})^{1/2}$. 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}.

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 and the logistic mixed-effects models. Results are based on* 1000 *independent replications*

Set-up				Means and ESD's		Sizes of z-test	
R_W	β_1	K	$V(\hat{\beta}_1)$	V_{LZ}	V_N	LZ	N
				Linear mixed-effects model			
I	0	10	0·0719	0·0570 (0·0389)	0·0634 (0·0307)	1·128 (0·185)	0·078 (0·131)
	0	40	0·0168	0·0160 (0·0056)	0·0166 (0·0037)	0·060 (0·116)	0·055 (0·105)
	1	10	0·0719	0·0570 (0·0389)	0·0634 (0·0307)		
	1	40	0·0168	0·0160 (0·0056)	0·0166 (0·0037)		
CS	0	10	0·0533	0·0389 (0·0251)	0·0434 (0·0189)	0·130 (0·200)	0·088 (0·152)
	0	40	0·0116	0·0108 (0·0035)	0·0111 (0·0022)	0·068 (0·114)	0·059 (0·109)
	1	10	0·0533	0·0389 (0·0251)	0·0434 (0·0189)		
	1	40	0·0116	0·0108 (0·0035)	0·0111 (0·0022)		
				Logistic mixed-effects model			
I	0	20	0·0814	0·0710 (0·0240)	0·0744 (0·0160)	0·068 (0·124)	0·049 (0·107)
	0	40	0·0375	0·0343 (0·0081)	0·0352 (0·0050)	0·068 (0·114)	0·058 (0·102)
	1	20	0·1346	0·1105 (0·0584)	0·1173 (0·0487)		
	1	40	0·0569	0·0526 (0·0164)	0·0541 (0·0120)		
CS	0	20	0·0788	0·0655 (0·0245)	0·0690 (0·0163)	0·068 (0·135)	0·048 (0·112)
	0	40	0·0356	0·0322 (0·0083)	0·0331 (0·0052)	0·066 (0·107)	0·052 (0·098)
	1	20	0·1315	0·1056 (0·0573)	0·1098 (0·0463)		
	1	40	0·0555	0·0505 (0·0161)	0·0515 (0·0120)		

CS, compound symmetry; ESD, estimated standard deviation; LZ, the Liang & Zeger method; N, new method.

Our second model 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, $y_{ij} \sim \text{Bi}(1, \mu_{ij})$, 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 the binary logistic model. We 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 1. The same conclusion can be drawn: our variance estimator of $\hat{\beta}_1$ performs better than the Liang–Zeger estimator. 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, that is $R_W = I$, the Liang–Zeger estimator is unchanged, and our estimator, obtained by setting the diagonal elements of the R_{Ui} equal to 1, is presented in Table 2. It appears that our estimator still has a slight edge.

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. For the proposed new variance estimator, it is assumed that each subject has its own correlation matrix. Results are based on* 1000 *independent replications*

Set-up				Means and ESD's		Sizes of z-test	
R_W	β_1	K	$V(\hat{\beta}_1)$	V_{LZ}	V_N	LZ	N
I	0	20	0·0814	0·0710 (0·0240)	0·0714 (0·0231)	0·068 (0·124)	0·058 (0·116)
	0	40	0·0375	0·0343 (0·0081)	0·0344 (0·0078)	0·068 (0·114)	0·063 (0·112)
	1	20	0·1346	0·1105 (0·0584)	0·1136 (0·0518)		
	1	40	0·0569	0·0526 (0·0164)	0·0538 (0·0155)		

ESD, estimated standard deviation; LZ, the Liang & Zeger method; N, new method.

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

$$\mathrm{vec}(M_{LZ}) = \sum_{i=1}^{K} \{(D_i' V_i^{-1}) \otimes (D_i' V_i^{-1})\} \,\mathrm{vec}(S_i S_i'),$$

we have

$$\mathrm{cov}\{\mathrm{vec}(M_{LZ})\} = \sum_{i=1}^{K} C_i \Omega_i C_i',$$

where $C_i = (D_i' V_i^{-1}) \otimes (D_i' V_i^{-1})$ and $\Omega_i = \mathrm{cov}\{\mathrm{vec}(S_i S_i')\}$. Similarly,

$$\mathrm{vec}(M_N) = \sum_{i=1}^{K} C_i \,\mathrm{vec}\left\{ A_i^{1/2} \sum_{j=1}^{K} A_j^{-1/2} S_j S_j' A_j^{-1/2} A_i^{1/2}/K \right\},$$

and hence

$$\mathrm{cov}\{\mathrm{vec}(M_N)\} = \sum_{i=1}^{K} C_i \left\{ F_i \sum_{j=1}^{K} \left(\frac{1}{K^2} F_j^{-1} \Omega_j F_j^{-1} \right) F_i \right\} C_i',$$

where $F_i = A_i^{1/2} \otimes A_i^{1/2}$. Thus

$$\mathrm{cov}\{\mathrm{vec}(M_{LZ})\} - \mathrm{cov}\{\mathrm{vec}(M_N)\} = \sum_{i=1}^{K} C_i \left(\Omega_i - F_i \frac{\sum_{j=1}^{K} F_j^{-1} \Omega_j F_j^{-1}}{K^2} F_i \right) C_i'.$$

Under suitable conditions, $\sum_{j=1}^{K} F_j^{-1} \Omega_j F_j^{-1}/K$ 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 Ω_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

AGRESTI, A. (1990). *Categorical Data Analysis*. New York: Wiley.
CROWDER, M. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* **82**, 407–10.
DIGGLE, P. J., LIANG, K.-Y. & ZEGER, S. L. (1994). *Analysis of Longitudinal Data*. New York: Oxford University Press.
JOHNSON, N. L. & KOTZ, S. (1970). *Distributions in Statistics, Continuous Univariate Distributions*, **2**. Boston: Houghton-Mifflin.
LIANG, K. Y. & ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.
MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.
PRENTICE, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–48.
ROTNITZKY, A. & JEWELL, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for clustered data. *Biometrika* **77**, 485–97.
VONESH, E. F. & CHINCHILLI, V. M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker.
WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61**, 439–47.
ZEGER, S. L. & LIANG, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–30.
ZEGER, S. L., LIANG, K. Y. & ALBERT, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049–60.

[*Received April* 2000. *Revised March* 2001]