Small-sample Performance of the Score Test in GEE

Xu Guo and Wei Pan

Division of Biostatistics

School of Public Health

University of Minnesota

A460 Mayo Building (MMC 303)

Minneapolis, MN 55455-0378, USA

Http: //www.biostat.umn.edu/~weip

Email: weip@biostat.umn.edu

Phone: (612)626-2705, Fax: (612)626-0660

May 2002

Small-sample Performance of the Score Test in GEE

Summary

The sandwich variance estimator in generalized estimating equations (GEE) may not perform well when the number of independent clusters is small. This could jeopardize the validity of the robust Wald test that is based on the use of the sandwich estimator by causing inflated Type I errors. Recently various modifications have been proposed to improve the small-sample performance of the Wald test. Here we take a different approach considering the generalized score test and its small-sample performance. In a simulation study, we compare it to the Wald test for correlated Bernoulli, Binomial and Poisson responses respectively. The score test has size close to the nominal level even when the number of clusters is as small as 10, whereas the Wald test has size that is multiple times of the nominal level. In addition to hypothesis testing, a 95% equal-tail confidence interval can be constructed from the score statistic. It has coverage probability close to 0.95, whereas the coverage probability of a 95% equal-tail confidence interval based on the Wald statistic is smaller than 0.95. More impressively, the score test also has some advantages when compared to other recent modifications to the Wald test, including a direct bias-correction of the sandwich estimator (Mancl and DeRouen, 2001) and an approximate t or F test (Pan and Wall, 2002).

Key words: Bias-correction; Correlated data; GEE; Sandwich variance estimator; Score test; Wald test.

1. Introduction

Correlated data often arise from biomedical research due to repeated measurements on the same individual or clustered sampling. For example, in the Guidelines for Urinary Incontinence Discussion and Evaluation (GUIDE) study, there were 137 patients from 38 practices. Baseline survey data were used to predict individual patient's response to the question whether they were bothered by accidental urine loss (Preisser and Qaqish, 1999). The measurements on the patients from the same practice may be correlated. Appropriate analyses need to take account of the potential correlation.

The generalized estimating equation (GEE) method (Liang and Zeger, 1986) has been widely used to model correlated data and draw proper statistical inference. When the number of independent clusters is sufficiently large, GEE method has some desirable properties, which have contributed to its popularity. The regression coefficent estimates from GEE are consistent and asymptotically normal. Their covariance is consistently estimated by the robust variance estimator of sandwich form (so-called sandwich estimator), which is robust to the misspecification of the covariance of the correlated responses.

But for small samples, the sandwich estimator does not perform well (Drum and McCullagh, 1993; Kauermann and Carroll, 2001), which was confirmed by some simulation studies (Emrich and Piedmonte, 1992; Gunsolley, Getchell and Chinchilli, 1995). This has jeopardized the validity of the robust Wald test by causing inflated Type I errors than specified nominal levels, and thus lower coverage probabilities of the corresponding confidence interval. How to improve the performance of the sandwich estimator and Wald test has become an active research topic; for a review on the current methods for small sample inference for correlated data, see Feng and Braun (2002). Most of the existing approaches focus on improving the small-sample property of the sandwich estimator to control the size of the resulting test (Pan, 2001; Daniels and Kass, 2001 and references below). In general,

there are two ways (Fay et al, 1998; Fay and Graubard, 2001). One is to correct the bias of the sandwich estimator or to reduce its variability, and the other is to take account of its variability. Since the sandwich estimator is downward biased for the covariance matrix of the regression coefficient estimates and the bias is especially large for small sample sizes, Mancl and DeRouen (2001) proposed to use a bias-corrected sandwich estimtor, instead of the usual sandwich estimator, to better control the size of the Wald test. An alternative was developed by Pan and Wall (2002) to take account of the variability of the sandwich estimator and construct an approximate t or F test, which is guaranteed to have a smaller size than the Wald test.

Here, we take a different approach. Instead of using the Wald test and trying to improve its performance, we consider the (generalized) score test for correlated data. Breslow (1990) derived a score test for independent response data, which is applicable to overdispersed quasi-likelihood models. It has two versions, an empirical and a model-based score tests, depending on using the sandwich or the model-based variance estimator. The empirical score test is robust to misspecification of the mean/variance relationship. It can achieve the nominal level for moderate sample sizes. Here we consider the score test for correlated data. We compared the proposed empirical score test to other score tests that appeared in the literature. We found that it is actually the same as the generalized score test given by Rotnitzky and Jewell (1990), though they look different. Since the latter authors did not provide details in their derivation, we give the sketch of ours following Breslow's idea in Appendix. We also found that the score test derived by Lefkopoulou and Ryan (1993) in a dose response experiment setting is a special case of our proposed score test. Boos (1992) also studied the (generalized) score test in a general setting.

Since the performance of the score test, especially for small samples, has not been investigated, even though it is an important issue in practice, we focus on this issue here.

As we will describe later, it turns out that the score test has surprisingly good small-sample perforamce, in contrast to its lack of use in many popular statistical packages that implement GEE, such as SPlus. In Sections 2 and 3, we briefly review the GEE method, the (robust) Wald test and a bias-corrected Wald test. In Section 4, we give the score test for correlated data and its bias-corrected version. In Section 5, we conduct simulations to evaluate its performance and compare it to the Wald test and to the approximate t or F test. We also consider constructing a confidence interval by inverting the score statistic. We illustrate our methods using the data from the GUIDE study in Section 6 and end with a short discussion in Section 7.

2. Generalized Estimating Equations

Suppose we have a correlated data set with K clusters. For each cluster i $(i = 1, ..., K)$, there are a response vector, $Y_i = (Y_{i1}, ..., Y_{in_i})'$, and an $n_i \times p$ matrix X_i of covariates associated with each resposne Y_i. Y_{ij}'s $(j = 1, ..., n_i)$ are assumed correlated within each cluster and independent across clusters. The marginal expectation of the response, $E(Y_{ij}) = \mu_{ij}$, is modelled by a regression equation $g(\mu_{ij}) = X_{ij}\beta$, where $\beta = (\beta_1, ...\beta_p)'$ is a p-dimensional vector of unknown regression coefficients and $g(\cdot)$ is a known link function. The marginal variance is $Var(Y_{ij}) = v(\mu_{ij})\phi$, where v is a known variance function and ϕ is a scale parameter. The within-cluster correlation matrix $R_0 = corr(Y_i)$ is generally unknown. We can consistently estimate β by specifying a working correlation matrix $R_W(\alpha)$, which may depend on some parameter α, and solving the following generalized estimating equations,

$$U(\beta, \phi) = \sum_{i=1}^{K} D_i' V_i^{-1} S_i = 0, \qquad (1)$$

where $D_i = \partial\mu_i/\partial\beta$, $\mu_i = (\mu_{i1}, ..., \mu_{in_i})'$, $V_i = \phi A_i^{1/2} R_W(\alpha) A_i^{1/2}$, $A_i = diag(v(\mu_{i1}), ..., v(\mu_{in_i}))$, $S_i = Y_i - \mu_i$.

Under mild regularity conditions, $\hat{\beta}$ is consitent and asymptotically normal (Liang and

Zeger, 1986). $Cov(\hat{\beta})$ can be consistently estimated by the so-called sandwich estimator,

$$V_S = (\sum_{i=1}^{K} D_i' V_i^{-1} D_i)^{-1} (\sum_{i=1}^{K} D_i' V_i^{-1} S_i S_i' V_i^{-1} D_i)(\sum_{i=1}^{K} D_i' V_i^{-1} D_i)^{-1}, \quad (2)$$

where β and α are replaced by their estimates $\hat{\beta}$ and $\hat{\alpha}$. If V_i correctly specifies $Cov(Y_i)$, the covariance matrix of $\hat{\beta}$ is consistently estimated by

$$V_M = (\sum_{i=1}^{K} D_i' V_i^{-1} D_i)^{-1},$$

which is called model-based or naive variance estimator.

3. Robust Wald Tests

To test a single regression coefficient β_k, $H_0 : \beta_k = 0$, we can use z statistic, $z = \hat{\beta}_k / \sqrt{V_{Sk}}$, where V_{Sk} is the kth diagonal element of V_S. It has an asymptotic standard normal distribution $N(0,1)$ under H_0. Obviously, a corresponding confidence interval for β_k can be constructed based on the z statistic. To test multiple regression coefficients, $H_0 : \beta = 0$, we can use a Wald chi-squared statistic, $W = \hat{\beta}' V_S^{-1} \hat{\beta}$. It has an asymptotic chi-squared distribtion χ_p^2 under H_0, where p is the dimension of β. The z test is just a special case of a Wald chi-squared test with $p = 1$. We call both as a robust Wald test. When the number of clusters is large, valid statistical inference can be accomplished by using the Wald test. Unfortunately, it does not work well for small sample sizes. Its test size is inflated and the corresponding confidence interval has lower coverage probabilities.

Since $E(S_i S_i') \doteq (I - H_i) Cov(Y_i)(I - H_i')$, Mancl and DeRouen (2001) proposed to correct the downward bias of the sandwich estimator V_S by replacing $S_i S_i'$ with $(I - H_i)^{-1} S_i S_i' (I - H_i')^{-1}$, where I is an $n_i \times n_i$ identity matrix and $H_i = D_i V_M D_i' V_i^{-1}$. The bias-corrected sandwich estimator was denoted as V_{BC},

$$V_{BC} = (\sum_{i=1}^{K} D_i' V_i^{-1} D_i)^{-1} (\sum_{i=1}^{K} D_i' V_i^{-1} (I - H_i)^{-1} S_i S_i' (I - H_i')^{-1} V_i^{-1} D_i)(\sum_{i=1}^{K} D_i' V_i^{-1} D_i)^{-1}. \quad (3)$$

4. Empirical and Model-based Score Tests

To test $H_0 : \beta_2 = \beta_2^0$ versus $H_1 : \beta_2 \neq \beta_2^0$, we decompose the p-dimensional regression coefficient vector β as $(\beta_1', \beta_2')'$, where β_1 and β_2 are vectors of dimensions of p_1 and p_2. The generalized estimating equation (1) can be decomposed accordingly as $U(\beta_1, \beta_2, \phi) = (U_1'(\beta_1, \beta_2, \phi), U_2'(\beta_1, \beta_2, \phi))'$. Under H_0, we can solve $U_1(\beta_1, \beta_2^0, \phi) = 0$ to get an estimate of β_1, denoted as $\hat{\beta}_1$. $U_2(\hat{\beta}_1, \beta_2^0, \phi)$ is used to test the hypothesis, $H_0 : \beta_2 = \beta_2^0$. A_{11}, A_{12}, A_{21}, A_{22}, B_{11}, B_{12}, B_{21} and B_{22} are the corresponding decomposed submatrices from $A = \lim_K K^{-1} A_K$, $B = \lim_K K^{-1} B_K$, where

$$A_K = -E\frac{\partial U}{\partial \beta'} = \sum_{i=1}^{K} \left(\frac{\partial \mu_i}{\partial \beta'}\right)' V_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta'}\right), \tag{4}$$

$$B_K = EUU' = \sum_{i=1}^{K} \left(\frac{\partial \mu_i}{\partial \beta'}\right)' V_i^{-1} Var(y_i) V_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta'}\right), \tag{5}$$

$$Var(y_i) = E(y_i - \mu_i)(y_i - \mu_i)'.$$

The variance of $(1/\sqrt{K})U_2(\hat{\beta}_1, \beta_2^0, \phi)$ is given by

$$Cov[(1/\sqrt{K})U_2(\hat{\beta}_1, \beta_2^0, \phi)] \doteq B_{22} - A_{21}A_{11}^{-1}B_{12} - B_{21}A_{11}^{-1}A_{12} + A_{21}A_{11}^{-1}B_{11}A_{11}^{-1}A_{12}, \tag{6}$$

which is robust to misspecification of the working correlation matrix $R_W(\alpha)$. If V_i is correctly specified, then $A = B$, and thus

$$Cov[(1/\sqrt{K})U_2(\hat{\beta}_1, \beta_2^0, \phi)] \doteq A_{22} - A_{21}A_{11}^{-1}A_{12}. \tag{7}$$

A covariance estimate $\hat{\Sigma}$ for $U_2(\hat{\beta}_1, \beta_2^0, \phi)$ is obtained by substituting a $\hat{\phi}$ estimated from the full model for ϕ and $(\hat{\beta}_1', \beta_2^{0\prime})'$ for β. The score statistic is constructed as

$$S = U_2(\hat{\beta}_1, \beta_2^0, \hat{\phi})'\hat{\Sigma}^{-1}U_2(\hat{\beta}_1, \beta_2^0, \hat{\phi}). \tag{8}$$

S is called the empirical or the model-based score statistic, depending on whether the empirical or the model-based variance estimate for $Cov[U_2(\hat{\beta}_1, \beta_2^0, \phi)]$ is used. The score statistic follows an asymptotic chi-squared distribution $\chi^2_{p_2}$ under H_0.

We have applied Mancl and DeRouen's bias correction method to the proposed score test. By using $(I-H_i)^{-1}(y_i-\hat{\mu}_i)(y_i-\hat{\mu}_i)'(I-H_i')^{-1}$ instead of $(y_i-\hat{\mu}_i)(y_i-\hat{\mu}_i)'$ to estimate $Var(y_i)$, we can get a bias-corrected empirical or model-based score test.

5. Simulation

5.1 Hypothesis testing

Simulations were conducted to study the performance of the score tests for correlated Bernoulli, Binomial, and Poisson responses respectively. Correlated responses can be easily generated from a generalized normal random-effects model, which can be well approximated by a corresponding generalized marginal model (Zeger, Liang and Albert, 1988). For each data configuration, 500 simulations were generated. For each simulation, the correct generalized marginal regression model was fitted using the independence working correlation structure in GEE. We have compared the score test to the Wald test for both individual and joint hypotheses. Actually, the proposed score test can be applied to correlated data with equal or unequal cluster size. But, for simplicity, simulations were run with an equal cluster size. Later, we demonstrate the score test for the data with unequal cluster sizes taken from the GUIDE study.

5.1.1 Bernoulli responses

Correlated Bernoulli responses, Y_{ij}'s, are randomly generated from $Bin(1,\mu_{ij})$, where μ_{ij} are generated from one of the following two logistic normal random-effect models,

$$logit(\mu_{ij}|b_i) = \beta_0 + x_{ij}\beta_1 + b_i, \tag{9}$$

$$logit(\mu_{ij}|b_i) = \beta_0 + x_{1ij}\beta_1 + x_{2ij}\beta_2 + x_{3ij}\beta_3 + b_i, \tag{10}$$

where $\beta_0 = 0$, $i = 1, ..., K$ $(K = 10, 20, 30)$, and $j = 1, ..., 20$. The covariates x_{ij}, x_{1ij}, x_{2ij} and x_{3ij} are all iid from $Bin(1, 1/2)$, b_i's are iid from $N(0, 1)$, and they are independent of each other. Only individual hypothesis test, $H_0 : \beta_1 = 0$, is conducted for simulated data

from (9), whereas both individual and joint hypothesis tests, $H_0 : \beta_1 = 0$ and $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, are conducted for simulated data from (10).

The sizes of the z test and of the score tests are evaluated as the observed fractions of the tests rejecting a null hypothesis when the null hypothesis is true. The sizes of individual hypothesis tests for data from (9) and their 95% equal-tail normal confidence intervals are shown in Table 1 for both 0.05 and 0.01 nominal levels. We have truncated the lower bound of a confidence interval at 0 if it is smaller than 0. When K is 10, the size of z test is 0.096, almost twice of the nominal level of 0.05, As K increases to 30, it gets closer and closer to the nominal levels. However, the sizes of both empirical and model-based score tests are much closer to the nominal levels. When K is 10, the sizes of the empirical and model-based score tests are 0.042 and 0.046 at the nominal level of 0.05. At the sample size of 30, all three tests have sizes close to the nomimal level of 0.05. We can compare the powers of the three tests. With different values of β_1, the power difference of the three tests is very small, at most 6% at the nominal level of 0.05, indicating there is no significant power loss using the score tests. Although the number of simulations (500) is too small to allow accurate estimation of rejection probabilities at the nominal level of 0.01, there is no obvious disagreement to the conclusion that the proposed score test has size close to nominal levels and maintains reasonble power in the mean time.

For the joint hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, the size of the Wald test is more inflated than that for individual hypothesis test, $H_0 : \beta_1 = 0$ (Table 2). When K is 10, the size of the Wald test is 0.240 at the nominal level of 0.05, whereas it is 0.028 for the empirical score test and 0.050 for the model-based score test. Even when K increases to 40, the Wald test still has inflated size of 0.084. For different simulation set-ups, the sizes of the score test are always much closer to the nominal levels.

It is interesting to find that the model-based score test performs very well, sometimes

even better than the empirical score test. It is mainly because the correlation introduced by normal random effect b_i is small. The true correlation matrix of the responses is very close to the indepedent working correlation matrix. In this situation, the empirical score test may be a little conservative since we lose some efficientcy by using sandwich estimator instead of the model-based one.

Table 1 about here

Table 2 about here

5.1.2 Binomial responses

Correlated Binomial responses, Y_{ij}'s, are randomly generated from $Bin(N_{ij}, p_{ij})$, where N_{ij} is an interger between 1 and 11 by sampling with replacement, and p_{ij} is generated from the following two logistic normal random-effect models,

$$logit(p_{ij}|b_i) = \beta_0 + x_{ij}\beta_1 + b_i, \tag{11}$$

$$logit(p_{ij}|b_i) = \beta_0 + x_{1ij}\beta_1 + x_{2ij}\beta_2 + x_{3ij}\beta_3 + b_i, \tag{12}$$

where $\beta_0 = 0$, $i = 1, ..., K$ $(K = 10, 20, 30, 40)$, $j = 1, ..., 4$. The covariates x_{ij}, x_{1ij}, x_{2ij} and x_{3ij} are all *iid* from a Bernoulli dstribution $Bin(1, 1/2)$, b_i's are *iid* from $N(0, 1)$, and they are independent of each other.

For simulated data from (11), the sizes of individual hypothesis tests at the nominal levels of 0.05 and 0.01 and their 95% normal confidence intervals are shown in Table 3. When K is 10, the size of z test is 0.110, more than twice of the nominal level of 0.05. Score tests perform much better. The size is 0.050 for the empirical score test, 0.074 for the model-based score test at the nominal level of 0.05. z test has inflated size (0.074) even when K is as large as 40. But the empirical score test performs very well under all the simualtion set-ups. Its size is very close to the nominal level of 0.05. For the simulated data from (12), both individual and joint hypothesis tests are conducted (Table 4). The Wald

test has a much highly inflated size for a joint hypothesis than for an individual hypothesis. The empirical score test is a little conservative for the joint hypothesis relative to that for the individual hypothesis. We notice that the model-based score test has inflated size, but still performs better than the Wald test. Maybe ignoring the within-cluster correlation by using independent working correlation matrix here has caused significant deviation from the true correlation structure of the responses. Also, we know that the model-based one does not model the overdispersed marginal variance correctly. The empirical score test has shown its advantage of robustness to the misspecification of the covariance structure of the responses over the model-based one.

Table 3 about here

Table 4 about here

5.1.3 Poisson responses

Correlated Poisson responses, Y_{ij}'s, are randomly generated from $Poisson(\mu_{ij})$, where μ_{ij} are generated from one of the following two normal random-effect poisson regression models,

$$log(\mu_{ij}|b_i) = \beta_0 + x_{ij}\beta_1 + b_i, \tag{13}$$

$$log(\mu_{ij}|b_i) = \beta_0 + x_{1ij}\beta_1 + x_{2ij}\beta_2 + x_{3ij}\beta_3 + b_i, \tag{14}$$

where $\beta_0 = 0$, $i = 1, ..., K(K = 10, 20, 30, 40)$, $j = 1, ..., 4$. The covariates x_{ij}, x_{1ij}, x_{2ij} and x_{3ij} are all iid from a Bernoulli distribution $Bin(1, 1/2)$, b_i's are iid from $N(0, 1)$, and they are independent of each other.

For the individual hypothesis $H_0 : \beta_1 = 0$ with data generated from (13), the size of z test (0.126) at the nominal level of 0.05 is dramatically inflated when K is 10 (Table 5). It is 0.020 for the empirical score test and 0.062 for the model-based score test. When K is 20, the size of z test is 0.112 at the nominal level of 0.05, whereas it is 0.044 for the empirical score test and 0.080 for the model-based score test. As sample size increases, the size of z

test is approaching the nominal level of 0.05. But, even when K is 40, its size (0.094) is still almost twice of the nominal level of 0.05. The size of the empirical score test is much closer to the nominal levels. But it may be a little conservative. The size of the model-based score test is inflated for correlated Poisson responses.

For the joint hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, the size of the Wald test is much more inflated, at most, 7-8 times of the nominal level of 0.05 (Table 6). When K is 40, it is still more than 3 times of the nominal level of 0.05. The empirical score test performs much better except that it is a little conservative.

Table 5 about here

Table 6 about here

5.1.4 A comparison with other approaches

Mancl and DeRouen (2001) found that a bias-correction method was helpful to improve the small-sample property of the sandwich estimator, but not enough to bring down the inflated test size to nominal level. Pan and Wall (2002) developed an alternative method that takes account of the variability of the sandwich estimator. To test an individual regression coefficient, say $H_0 : \beta_k = 0$, they constructed a t statistic, which had an approximate t distribution with finite degrees of freedom under null hypothesis. Similarly, they constructed a Wald statistic for a joint hypothesis test, say $H_0 : \beta = 0$, and they used a scaled F distribution as the null distribution of the Wald statistic. Through simulations, they found that their proposed t- or F-test is competitive when compared with other approaches. In particular, they found that an additional bias correction on the sandwich estimator was not necessary in the corresponding F-test.

Since our simulation set-ups are exactly the same as those in Pan and Wall's paper, we can compare our score test to their t or F test directly. First, we consider testing for an individual hypothesis test, $H_0 : \beta_1 = 0$. For correlated Bernoulli responses with sample

size of 10, the size is 0.042 for the empirical score test, 0.046 for the model-based score test at the nominal level of 0.05, whereas it is 0.074 for the t test; for correlated Binomial responses with sample size of 10, the size is 0.050 for the empirical score test, 0.074 for the model-based score test at the nominal level of 0.05, whereas it is 0.076 for the t test; for correlated Poisson responses with sample size of 10, the size is 0.020 for the empirical score test, 0.062 for the model-based score test at the nominal level of 0.05, whereas it is 0.096 for the t test. So, the empirical score test performs better than the t test. We also have compared the score test with the F test for a joint hypothesis, $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. The empirical score test is a little conservative, but the F test turns out to be even more conservative. So, in general, the empirical score test performs better than the t or F test.

It is natural to apply Mancl and DeRouen's bias correction approach to our score test. We implemented it and it did not improve the perfomance of the empirical score test any further. The empirical score test with usual sandwich estimator is enough to bring the test size near the nominal level and additional bias-correction is actually not necessary.

5.2 Interval estimation

Confidence intervals for the regression coefficients can be easily constructed from the Wald statistic. Unfortunately, they have lower coverage probability than the nominal level for small samples. By inverting the proposed score test, we can construct the corresponding confidence interval. Since the confidence interval based on the Wald statistic is often too narrow, we enlarge it generously by taking the multiples of the standard error of the estimated regression coefficient to obtain a reasonable starting point to find the score-test-based confidence interval. We then use the bisection algorithm (Press et al 1992, p.353) to search within this enlarged interval for the two endpoints of of the confidence interval based on the score statistic. We compared 95% equal-tail confidence intervals based on the Wald statistic and on the empirical score statistic for correlated Bernoulli responses generated from

(9). For each data configuration, 500 simulations were run to obtain the average length of the intervals and their coverage probability (Table 7). When the true value of β_1 is zero, the coverage probabilities are consistent with the test sizes for the individual hypothesis $H_0 : \beta_1 = 0$ given in Table 1. When the true value of β_1 is 0.5, the coverage probability for the Wald confidence interval is smaller than the specified confidence level 0.95, whereas it is close to 0.95 for the confidence interval based on the empirical score test. It seems that the 95% equal-tail confidence intervals based on the Wald statistic is narrower than that based on the empirical score statistic. By using the empirical score test, we have widened the length of the interval appropriately to accomplish correct coverage probability. The empirical score test can not only control the size in hypothesis testing effectively, but also provide confidence intervals with correct coverage probability.

Table 7 about here

6. Example

We illustrate the application of the empirical score test using the data taken from the GUIDE study. Preisser and Qaqish (1999) used five covariates measured from a baseline survey to predict individual patient's response to the question whether they were bothered by accidental loss of urine. Here, we fit a logistic regression model including the five covariates, $female$, age, $dayacc$, $severe$ and $toilet$. The Wald test and the empirical score test are conducted to study whether the five covariates have any effect on the response individually or jointly. The estimated regression coefficients, P-values and 95% equal-tail confidence intervals from the Wald test and empirical score test are summarized in Table 8. There is not much difference in the P-values from the Wald and empirical score tests since the sample size is not very small. The individual effects of both $Dayacc$ and $Severe$ are statistically significant. For $Severe$, the P-value from the empirical score test (0.042) is larger than that

from the Wald test (0.024), indicating we get more significant testing results from the Wald test. This is consistent with the results from simulations. Since the Wald test has inflated Type I error when sample size is small, we should be cautious about significant results from the Wald test. As for confidence intervals, the 95% equal-tail confidence intervals from the empirical score statistic is wider than those from the Wald statistic.

Table 8 about here

7. Discussion

The robust Wald test is constructed from the comparison between regression coefficient estimates and their robust covariance estimate. It performs well when the sample size is sufficiently large. Due to its easy access in SAS or Splus statistical packages, it has been widely, and indeed almost exclusively, used to conduct hypothesis testing and construct confidence intervals. Even with large samples, one of its drawbacks is that it is not invariant under reparameterization (Vaeth, 1985) whereas the score test is invariant. More importantly, when the sample size is small, the robust Wald test does not perform well. Our simulation studies have confirmed that it has dramatically inflated Type I errors and the corresponding confidence interval has a lower coverage probability than a specified confidence level. In this paper, we considered the (generalized) score test to analyze correlated data. It is easy to conduct and has very impressive performance. It has Type I error closer to the nominal level, and is even more effective than the approximate t or F test to control the inflated test size. It can also provide confidence intervals with effective coverage probability. Thus, the score test is better than the Wald test for hypothesis testing and interval estimation for correlated data of small sample size. When the true within-cluster correlation structure is unknown, the empirical score test is recommended because it is robust to the misspecification of the working covariance structure.

APPENDIX

Derivation of the score test

$U_2(\hat{\beta}_1, \beta_2^0, \phi)$ is used to test the hypothesis, $H_0 : \beta_2 = \beta_2^0$. By using first-order Taylor expansion, we get

$$0 = (1/\sqrt{K})U_1(\hat{\beta}_1, \beta_2^0, \phi) \doteq (1/\sqrt{K})\{U_1(\beta_1, \beta_2^0, \phi) + \frac{\partial U_1}{\partial \beta_1'}(\hat{\beta}_1 - \beta_1)\}$$

$$(1/\sqrt{K})U_2(\hat{\beta}_1, \beta_2^0, \phi) \doteq (1/\sqrt{K})\{U_2(\beta_1, \beta_2^0, \phi) + \frac{\partial U_2}{\partial \beta_1'}(\hat{\beta}_1 - \beta_1)\}$$

From the first equation, we get

$$\hat{\beta}_1 - \beta_1 \doteq -\left(\frac{\partial U_1}{\partial \beta_1'}\right)^{-1} U_1(\beta_1, \beta_2^0, \phi).$$

Plug this into the second equation, we get

$$(1/\sqrt{K})U_2(\hat{\beta}_1, \beta_2^0, \phi) \doteq (1/\sqrt{K})\{U_2(\beta_1, \beta_2^0, \phi) - \frac{\partial U_2}{\partial \beta_1'}\left(\frac{\partial U_1}{\partial \beta_1'}\right)^{-1} U_1(\beta_1, \beta_2^0, \phi)\}.$$

It follows an asymptotic multivariate normal distribution with zero mean and variance

$$\begin{aligned}
& Cov[(1/\sqrt{K})U_2(\hat{\beta}_1, \beta_2^0, \phi)] \\
\doteq\ & Cov\left[(1/\sqrt{K})\left(-\frac{\partial U_2}{\partial \beta_1'}\left(\frac{\partial U_1}{\partial \beta_1'}\right)^{-1}, I\right)\begin{pmatrix} U_1(\beta_1, \beta_2^0, \phi) \\ U_2(\beta_1, \beta_2^0, \phi) \end{pmatrix}\right] \\
=\ & \frac{1}{K}\left(-\frac{\partial U_2}{\partial \beta_1'}\left(\frac{\partial U_1}{\partial \beta_1'}\right)^{-1}, I\right) Cov\begin{bmatrix} U_1(\beta_1, \beta_2^0, \phi) \\ U_2(\beta_1, \beta_2^0, \phi) \end{bmatrix}\left(-\frac{\partial U_2}{\partial \beta_1'}\left(\frac{\partial U_1}{\partial \beta_1'}\right)^{-1}, I\right)' \\
\doteq\ & (-A_{21}A_{11}^{-1}, I)\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}\begin{pmatrix} (-A_{21}A_{11}^{-1})' \\ I \end{pmatrix} \\
=\ & B_{22} - A_{21}A_{11}^{-1}B_{12} - B_{21}A_{11}^{-1}A_{12} + A_{21}A_{11}^{-1}B_{11}A_{11}^{-1}A_{12},
\end{aligned}$$

where I is a $p_2 \times p_2$ identity matrix. We can construct a score statistic based on the above results.

References

Boos, D.D. (1992). On generalized score tests. *The American Statistician* **46**, 327-333.

Breslow, N. (1990). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association* **85**, 565-571.

Daniels, M.J. and Kass, R.E. (2001). Shrinkage estimators for covariance matrices. *Biometrics* **57**, 1173-1184.

Drum, M., McCullagh, P. (1993). Comment. *Statistical Science* **8**, 300-301.

Emrich, L.J., Piedmonte, M.R. (1992). On some small sample properties of generalized estimating equation estimates for multivariate dichotomous outcomes. *Journal of Statistical Computation and Simulation* **41**, 19-29.

Fay, M.P. and Graubard B.I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* **57**, 1198-1206.

Fay, M.P., Graubard B.I., Freedman L.S., Midthune D.N. (1998). Conditional logistic regression with sandwich estimators: application to meta analysis. *Biometrics* **54**, 195-208.

Feng, Z.D., Braun T.M. (2002). Small sample inference for clustered data. *To appear in* .

Gunsolley, J.C., Getchell, C., Chinchilli, V.M. (1995). Small sample characteristics of generalized estimating equations. *Communications in Statistics - Simulation* **24**, 869-878.

Kauermann, G. and Carroll, R.J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *JASA* **96**, 1387-1396.

Lefkopoulou, M. and Ryan, L. (1993). Global tests for multiple binary outcomes. *Biometrics* **49**, 975-988.

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

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

Pan, W. (2001). On the robust variance estimator in generalised estimating equations. *Biometrika* **88**, 901-906.

Pan, W. and Wall, M. (2002). Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine* **21**, 1429-1441.

Preisser, J.S. and Qaqish, B.F. (1999). Robust regression for clustered data with application to binary responses. *Biometrics* **55**, 574-579.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992). *Numerical Recipes in C, The Art of Scientific Computing*. 2nd ed. Cambridge: New York.

Rotnitzky, A. and Jewell, N.P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster data. *Biometrika* **77**, 485-489.

Vaeth, M. (1985). "Wald's test in exponential families". (1985) *International Statistical Review* **53**, 199-214.

Vonesh, E.F. and Chinchilli, V.M. (1997). Linear and nonlinear models for the analysis of repeated measurements. Marcel Dekker: New York.

Zeger, S.L., Liang K.Y. and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049-60.

Table 1: Empirical size and power (and their 95% confidence interval of the α-level z-test and score test for testing $H_0 : \beta_1 = 0$ in a mixed-effects logistic regression model for correlated Bernoulli responses.

Set-up		Wald-test		Empirical score test		Model-based score test	
β_1	K	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
0	10	.096	.034	.042	.004	.046	.006
		(.070, .122)	(.018, .050)	(.024, .060)	(.000, .010)	(.028, .064)	(.000, .013)
0	20	.054	.024	.034	.008	.040	.006
		(.034, .074)	(.011, .037)	(.018, .050)	(.000, .016)	(.023, .057)	(.000, 0.013)
0	30	.056	.008	.046	.002	.050	.002
		(.036, .076)	(.000, 0.016)	(.028, .064)	(.000, .006)	(.031, .069)	(.000, .006)
0.4	30	.512	.284	.468	.220	.490	.268
		(.468, .556)	(.244, .324)	(.424, .512)	(.184, .256)	(.446, .534)	(.229, .307)
0.5	30	.704	.480	.650	.358	.692	.444
		(.664, .744)	(.436, .524)	(.608, .692)	(.316, .400)	(.652, .732)	(.400, .488)
0.6	30	.844	.650	.814	.540	.858	.676
		(.812, .876)	(.608, .692)	(.780, .848)	(.496, .584)	(.827, .889)	(.635, .717)
0.7	30	.942	.818	.916	.730	.928	.816
		(.922, .962)	(.784, .852)	(.892, .940)	(.691, .769)	(.905, .951)	(.782, .850)

Table 2: Empirical size (and its 95% confidence interval) of the α-level Wald χ^2 and score tests for testing $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ or $H_0 : \beta_1 = 0$ in a mixed-effects logistic regression model for correlated Bernoulli responses.

Set-up		Wald-test		Empirical score test		Model-based score test	
H_0	K	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
$\beta_1 = \beta_2 = \beta_3 = 0$	10	.240	.136	.028	.000	.050	.014
		(.203, .277)	(.106, .166)	(.014, .042)	(.000, .000)	(.031, .069)	(.004, .024)
$\beta_1 = \beta_2 = \beta_3 = 0$	20	.120	.050	.034	.000	.048	.004
		(.092, .148)	(.031, .069)	(.018, .050)	(.000, .000)	(.029, .067)	(.000, .010)
$\beta_1 = \beta_2 = \beta_3 = 0$	30	.096	.026	.032	.002	.038	.006
		(.070, .122)	(.012, .040)	(.017, .047)	(.000, .006)	(.021, .055)	(.000, .013)
$\beta_1 = \beta_2 = \beta_3 = 0$	40	.084	.024	.048	.004	.056	.006
		(.060, .108)	(.011, .037)	(.029, .067)	(.000, .010)	(.036, .076)	(.000, .013)
$\beta_1 = 0$	10	.100	.032	.042	.004	.058	.014
		(.074, .126)	(.017, .047)	(.024, .060)	(.000, .010)	(.038, .078)	(.004, .024)
$\beta_1 = 0$	20	.066	.010	.032	.002	.040	.004
		(.044, .088)	(.001, .019)	(.017, .047)	(.000, .006)	(.023, .057)	(.000, .010)
$\beta_1 = 0$	30	.062	.016	.044	.006	.028	.006
		(.041, .083)	(.005, .027)	(.026, .062)	(.000, .013)	(.014, .042)	(.000, .013)
$\beta_1 = 0$	40	.068	.026	.056	.010	.062	.012
		(.046, .090)	(.012, .040)	(.036, .076)	(.001, .019)	(.041, .083)	(.002, .022)

Table 3: Empirical size (and its 95% confidence interval of the α-level z-test and score test for testing $H_0 : \beta_1 = 0$ in a mixed-effects logistic regression model for correlated Binomial responses.

Set-up		Wald-test		Empirical score test		Model-based score test	
β_1	K	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
0	10	.110	.034	.050	.000	.074	.020
		(.083, .137)	(.018, .050)	(.031, .069)	(.000, .000)	(.051, .097)	(.008, .032)
0	20	.090	.016	.056	.004	.088	.016
		(.065, .115)	(.005, .027)	(.036, .076)	(.000, .010)	(.063, .113)	(.005, .027)
0	30	.068	.016	.054	.010	.076	.026
		(.046, .090)	(.005, .027)	(.034, .074)	(.001, .019)	(.053, .099)	(.012, .040)
0	40	.074	.018	.060	.012	.092	.026
		(.051, .097)	(.006, .030)	(.039, .081)	(.002, .022)	(.067, .117)	(.012, .040)

Table 4: Empirical size (and its 95% confidence interval) of the α-level Wald χ^2 and score tests for testing $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ or $H_0 : \beta_1 = 0$ in a mixed-effects logistic regression model for correlated Binomial responses.

Set-up		Wald-test		Empirical score test		Model-based score test	
H_0	K	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
$\beta_1 = \beta_2 = \beta_3 = 0$	10	.208	.156	.008	.000	.090	.022
		(.172, .244)	(.124, .188)	(.000, .016)	(.000, .000)	(.065, .115)	(.009, .035)
$\beta_1 = \beta_2 = \beta_3 = 0$	20	.138	.044	.016	.000	.094	.024
		(.108, .168)	(.026, .062)	(.005, .027)	(.000, .000)	(.068, .120)	(.011, .037)
$\beta_1 = \beta_2 = \beta_3 = 0$	30	.110	.044	.040	.006	.090	.034
		(.083, .137)	(.026, .062)	(.023, .057)	(.000, .013)	(.065, .115)	(.018, .050)
$\beta_1 = \beta_2 = \beta_3 = 0$	40	.094	.038	.048	.004	.106	.028
		(.068, .120)	(.021, .055)	(.029, .067)	(.000, .010)	(.079, .133)	(.028, .042)
$\beta_1 = 0$	10	.142	.046	.062	.002	.094	.024
		(.111, .173)	(.028, .064)	(.041, .083)	(.000, .006)	(.068, .120)	(.011, .037)
$\beta_1 = 0$	20	.074	.016	.036	.004	.062	.016
		(.051, .097)	(.005, .027)	(.020, .052)	(.000, .010)	(.041, .083)	(.005, .027)
$\beta_1 = 0$	30	.062	.018	.052	.006	.062	.026
		(.041, .083)	(.006, .030)	(.033, .071)	(.000, .013)	(.041, .083)	(.012, .040)
$\beta_1 = 0$	40	.080	.030	.070	.018	.084	.032
		(.056, .104)	(.015, .045)	(.048, .092)	(.006, .030)	(.060, .108)	(.017, .047)

Table 5: Empirical size (and its 95% confidence interval of the α-level z test and score test for testing $H_0 : \beta_1 = 0$ in a mixed-effects Poisson regression model for correlated Poisson responses.

Set-up		Wald-test		Empirical score test		Model-based score test	
β_1	K	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
0	10	.126	.058	.020	.000	.062	.002
		(.097, .155)	(.038, .078)	(.008, .032)	(.000, .000)	(.041, .083)	(.000, .006)
0	20	.112	.036	.044	.002	.080	.020
		(.084, .140)	(.020, .052)	(.026, .062)	(.000, .006)	(.056, .104)	(.008, .032)
0	30	.094	.032	.028	.000	.060	.014
		(.068, .120)	(.017, .047)	(.014, .042)	(.000, .000)	(.039, .081)	(.004, .024)
0	40	.068	.014	.028	.004	.048	.012
		(.046, .090)	(.004, .024)	(.014, .042)	(.000, .010)	(.029, .067)	(.002, .022)

Table 6: Empirical size (and its 95% confidence interval) of the α-level Wald χ^2 and score tests for testing $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ or $H_0 : \beta_1 = 0$ in a mixed-effects Poisson regression model for correlated Poisson responses.

Set-up		Wald-test		Empirical score test		Model-based score test	
H_0	K	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
$\beta_1 = \beta_2 = \beta_3 = 0$	10	.374	.228	.006	.000	.102	.020
		(.332, .416)	(.191, .265)	(.000, .013)	(.000, .000)	(.075, .129)	(.008, .032)
$\beta_1 = \beta_2 = \beta_3 = 0$	20	.240	.126	.016	.000	.088	.022
		(.203, .277)	(.097, .155)	(.005, .027)	(.000, .000)	(.063, .113)	(.009, .035)
$\beta_1 = \beta_2 = \beta_3 = 0$	30	.154	.072	.026	.002	.076	.024
		(.122, .186)	(.049, .095)	(.012, .040)	(.000, .006)	(.053, .099)	(.011, .037)
$\beta_1 = \beta_2 = \beta_3 = 0$	40	.168	.060	.018	.000	.074	.018
		(.135, .201)	(.039, .081)	(.006, .030)	(.000, .000)	(.051, .097)	(.006, .030)
$\beta_1 = 0$	10	.150	.054	.028	.000	.072	.016
		(.119, .181)	(.034, .074)	(.014, .042)	(.000, .000)	(.049, .095)	(.005, .027)
$\beta_1 = 0$	20	.100	.042	.034	.000	.080	.016
		(.074, .126)	(.024, .060)	(.018, .050)	(.000, .000)	(.056, .104)	(.005, .027)
$\beta_1 = 0$	30	.104	.024	.048	.002	.072	.016
		(.077, .131)	(.011, .037)	(.029, .067)	(.000, .006)	(.049, .095)	(.005, .027)
$\beta_1 = 0$	40	.072	.028	.034	.002	.052	.024
		(.0490, .095)	(.014, .042)	(.018, .050)	(.000, .006)	(.033, .071)	(.011, .037)

Table 7: Average length and coverage probability of 95% equal-tail confidence intervals based on z statistic and empirical score statistic for correlated Bernoulli responses.

Set-up		z test				Empirical score test			
β_1	K	Average lower bound	Average upper bound	Average length	Coverage probability	Average lower bound	Average upper bound	Average length	Coverage probability
0	10	-.522	.513	1.035	.904	-.676	.668	1.344	.958
0	20	-.361	.389	.750	.946	-.406	.433	.839	.966
0	30	-.314	.311	.624	.944	-.337	.334	.671	.954
0.5	10	-.107	.939	1.046	.890	-.250	1.115	1.365	.950
0.5	20	.050	.813	.763	.928	.009	.865	.855	.944
0.5	30	.096	.728	.633	.890	.075	.756	.680	.930

Table 8: P-values of the various tests and 95% equal-tail confidence intervals of the regression coefficients for the GUIDE data.

Tested covariate	Coefficient estimate	Wald test		Empirical score test	
		P-value	CI	P-value	CI
Female	-.672	.281	(-1.893, .549)	.342	(-1.936, 1.262)
Age	-.641	.264	(-1.766, .484)	.330	(-1.901, .737)
Dayacc	.415	.000	(.223, .608)	.000	(.249, .705)
Severe	.829	.024	(.110, 1.547)	.042	(.041, 1.891)
Toilet	.111	.271	(-.087, .308)	.170	(-.048, .456)
All five		.000		.002	
Feamle+Age+Toilet		.460		.340	