STATISTICS IN MEDICINE
*Statist. Med.* 2002; **21**:1429–1441 (DOI: 10.1002/sim.1142)

# Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations

Wei Pan*,† and Melanie M. Wall

*Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455-0378, U.S.A.*

SUMMARY

The generalized estimating equation (GEE) approach is widely used in regression analyses with correlated response data. Under mild conditions, the resulting regression coefficient estimator is consistent and asymptotically normal with its variance being consistently estimated by the so-called sandwich estimator. Statistical inference is thus accomplished by using the asymptotic Wald chi-squared test. However, it has been noted in the literature that for small samples the sandwich estimator may not perform well and may lead to much inflated type I errors for the Wald chi-squared test. Here we propose using an approximate $t$- or $F$-test that takes account of the variability of the sandwich estimator. The level of type I error of the proposed $t$- or $F$-test is guaranteed to be no larger than that of the Wald chi-squared test. The satisfactory performance of the proposed new tests is confirmed in a simulation study. Our proposal also has some advantages when compared with other new approaches based on direct modifications of the sandwich estimator, including the one that corrects the downward bias of the sandwich estimator. In addition to hypothesis testing, our result has a clear implication on constructing Wald-type confidence intervals or regions. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: bias correction; $F$-test; GEE; robust variance estimator; $t$-test; Wald chi-squared test; $z$-test

## 1. INTRODUCTION

The marginal regression and its associated generalized estimating equation (GEE) approach have become one of the most widely used methods in dealing with correlated response data [1, 2]. In general, with a correlated data set, for each cluster or subject $i$ we have several measurements of a response $Y_{ij}$ and a $p$-dimensional covariate vector $X_{ij}$, $j=1,\ldots,n_i$ and $i=1,\ldots,K$. For simplicity we assume $n_i=n$ as in reference [1]. Denote $Y_i=(Y_{i1},\ldots,Y_{in})'$ and $X_i=(X_{i1}',\ldots,X_{in}')'$. The marginal regression model specifies a relation between the marginal

*Correspondence to: Wei Pan, Division of Biostatistics, A460 Mayo Building, MMC 303, Minneapolis, MN 55455-0378, U.S.A.
†E-mail: weip@biostat.umn.edu

Contract/grant sponsor: National Institute of Health; contract/grant number: R01 HL65462

*Received June 2001*
*Accepted August 2001*

mean $E(Y_{ij})=\mu_{ij}$ and the covariate $X_{ij}$ through a generalized linear model (GLM): (i) $g(\mu_{ij})=X_{ij}\beta$, where $\beta$ is an unknown $p$-dimensional vector of regression coefficients to be estimated, and $g$ is a known link function; (ii) the marginal variance is $\text{var}(Y_{ij})=v(\mu_{ij})\phi$, where $v$ is a known function and $\phi$ is a scale parameter which may need to be estimated; and (iii) the within-subject correlation matrix $\text{corr}(Y_i)$ is $R_0$ whose structure is in general unknown. It is assumed throughout that $Y_i$ and $Y_k$ are independent for any $i\neq k$. An attractive point of the GEE approach is that obtaining a consistent estimator for $\beta$ does not require $R_0$ to be specified correctly; instead, one can use some *working* correlation matrix $R_{\text{W}}(\alpha)$, which may depend on some parameter $\alpha$. Likewise, it is not necessary in this approach to specify $\text{var}(Y_{ij})$ correctly. Hence the GEE approach is non-likelihood based; the (asymptotic) validity of the GEE estimates only depends on the correct specification of the mean function of $Y_{ij}$ in (i). Denote $A_i=\text{diag}(v(\mu_{i1}),\ldots,v(\mu_{in}))$ and the *working* covariance matrix $V_i=\phi A_i^{1/2}R_{\text{W}}(\alpha)A_i^{1/2}$. Then the GEE approach estimates $\beta$ by solving the following estimating equations:

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

where $D_i=\partial\mu_i/\partial\beta'$, $S_i=Y_i-\mu_i$ and $\mu_i=(\mu_{i1},\ldots,\mu_{in})'$. Provided that we have a $K^{1/2}$-consistent estimator $\hat{\alpha}$, under mild regularity conditions, Liang and Zeger showed 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 (co)variance estimator:

$$V_{\text{S}}=\left(\sum_{i=1}^{K} D_i' V_i^{-1} D_i\right)^{-1}\left\{\sum_{i=1}^{K} D_i' V_i^{-1} S_i S_i' V_i^{-1} D_i\right\}\left(\sum_{i=1}^{K} D_i' V_i^{-1} D_i\right)^{-1} \tag{2}$$

where $\beta$ and $\alpha$ are replaced by their estimates $\hat{\beta}$ and $\hat{\alpha}$. If all the modelling assumptions (i)–(iii) are correct, then one can also use the usual model-based variance estimator

$$V_{\text{M}}=\left(\sum_{i=1}^{K} D_i' V_i^{-1} D_i\right)^{-1}$$

Now suppose that we want to test one of the regression coefficients, say $\beta_k$. Without loss of generality, the null hypothesis can be formulated as $H_0:\beta_k=0$. Then we can use the usual $z$-statistic $z=\hat{\beta}_k/\sqrt{V_{\text{S}k}}$, where $V_{\text{S}k}$ is the $k$th diagonal element of $V_{\text{S}}$. Under $H_0$, $z$ has an asymptotically normal distribution and a $p$-value can be thus obtained. Obviously, a corresponding confidence interval for $\beta_k$ can be also constructed based on the $z$-statistic. Bearing this in mind we only consider hypothesis testing in this paper. To test multiple parameters, without loss of generality, we consider the null hypothesis $H_0:\beta=0$, and Wald's chi-squared test can be applied; $W=\hat{\beta}' V_{\text{S}}^{-1}\hat{\beta}$ asymptotically has a chi-squared distribution $\chi_p^2$ under $H_0$, where $p$ is the dimension of $\beta$. The normal $z$-test is a special case of Wald's chi-squared test with $p=1$. Note that Wald's test is based on approximating $\text{cov}(\hat{\beta})$ by its consistent estimate $V_{\text{S}}$; in other words, both the bias and variation of $V_{\text{S}}$ are ignored.

In spite of the wide use of the sandwich estimator, it has been noted in the literature that the sandwich estimator may not work well with small samples (for example, Drum and McCullagh [3]). This point has been verified by other authors for binary data through simulation studies [4, 5]. Fay *et al.* [6] proposed an approximate $t$-test when using the sandwich estimator

 

in a different context. Here, following the same line, we propose a more general approach in using an approximate $t$- or $F$-test to test one or more than one parameter, rather than the aforementioned $z$-test or Wald chi-squared test, to take account of the variability of the sandwich estimator. The size of type I error of our proposed $t$- or $F$-test is guaranteed to be no larger than that of the $z$-test or Wald chi-squared test.

Furthermore, as the sample size increases, the size of type I error of our proposed $t$- or $F$-test approaches that of the $z$-test or Wald chi-squared test. We compare our approach with two other proposals based on direct modifications of the sandwich estimator [7, 8], the first of which corrects the bias of the sandwich estimator and the second is based on estimating the common correlation matrix (if it exists). We also explore a combined approach to account for both the variability and bias of the sandwich estimator, which however does not seem to perform better than adjusting for the variability alone.

## 2. VARIATION DUE TO THE SANDWICH ESTIMATOR

The basic idea of our proposal is to take account of some variability due to the sandwich variance estimator when testing regression coefficients. The working idea parallels that in testing the mean of a normal distribution with an unknown variance, where it is well-known that a $t$-test is preferred over a $z$-test.

In GEE, it has been observed that the model-based variance estimator $V_{\mathrm{M}}$ generally works well under correct modelling assumptions. Hence, for simplicity, we will ignore the variability in $D_i$ and $V_i$. This implies that we will treat $D_i$, $V_i$ and $V_{\mathrm{M}}$ as fixed (that is, non-random) and only estimate the variance of the middle piece of $V_{\mathrm{S}}$ in (2).

We first need to define an operator vec(.). For any matrix $U$, vec($U$) is a vector formed by stacking the columns of $U$ below one another. We want to derive an estimator for the covariance of vec$(\sum_{i=1}^{K} D_i' V_i^{-1} S_i S_i' V_i^{-1} D_i)$. Denote

$$P_i = \mathrm{vec}(D_i' V_i^{-1} S_i S_i' V_i^{-1} D_i)$$

Suppose that the mean vector and covariance matrix of $\sum_{i=1}^{K} P_i/K$ are, respectively, $Q$ and $T$. $\sum_{i=1}^{K} P_i/K$ itself is an unbiased estimator of $Q$, and $T$ can be estimated by the empirical covariance estimator

$$\hat{T} = \sum_{i=1}^{K} (P_i - \bar{P})(P_i - \bar{P})'/K(K-1)$$

where $\bar{P} = \sum_{i=1}^{K} P_i/K$.

Now we define the Kronecker product $\otimes$: for an $r \times s$ matrix $A = (a_{ij})$ and any matrix $B$

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1s}B \\ a_{21}B & a_{22}B & \dots & a_{2s}B \\ \dots & & & \\ a_{r1}B & a_{r2}B & \dots & a_{rs}B \end{pmatrix}$$

The covariance matrix of vec$(V_{\mathrm{S}}) = (V_{\mathrm{M}} \otimes V_{\mathrm{M}}) \sum_{i=1}^{K} P_i$ can be estimated using a formula for the $\otimes$ operation (Vonesh and Chinchilli, reference [15], p. 12) as

$$\widehat{\mathrm{cov}}(\mathrm{vec}(V_{\mathrm{S}})) = K^2 (V_{\mathrm{M}} \otimes V_{\mathrm{M}}) \hat{T} (V_{\mathrm{M}} \otimes V_{\mathrm{M}}) \tag{3}$$

 

## 3. APPROXIMATE $t$-TEST

### 3.1. Derivation

We first consider the situation of testing only one parameter with $H_0:\beta_k=0$. From $\widehat{\mathrm{cov}}(\mathrm{vec}(V_S))$ we can obtain the estimated variance of $V_{Sk}$. Now denote the (estimated) mean and variance of $V_{Sk}$ as $\sigma_k$ and $\tau_k$, respectively. We use a scaled chi-squared distribution $c\chi^2_d$ to approximate the distribution of $V_{Sk}$ such that $c$ and $d$ are determined by matching their first two moments [9] (see also, for example, Cox and Hinkley, reference [10], p. 463), and we obtain

$$c=\tau_k/2\sigma_k, \quad d=2\sigma_k^2/\tau_k$$

Since under $H_0$, $\hat{\beta}_k/\sqrt{\sigma_k}$ is approximately distributed $\mathrm{N}(0,1)$, and $V_{Sk}/c$ is approximately as $\chi^2_d$, then

$$t=\frac{\hat{\beta}_k/\sqrt{\sigma_k}}{\sqrt{V_{Sk}/cd}}=\frac{\hat{\beta}_k}{\sqrt{V_{Sk}}}$$

can be approximated by a $t$-distribution with degrees of freedom $d=2\sigma_k^2/\tau_k\approx 2V_{Sk}^2/\tau_k$, from which we can derive a $p$-value. Note that the test statistic for our $t$-test is exactly the same as that for the usual $z$-test, but rather than using the standard normal $\mathrm{N}(0,1)$, we use $t_d$ as the reference distribution. Also note that $\tau_k$ and thus $d$ can be estimated as we described earlier. In addition, if the variability of the sandwich estimator $V_{Sk}$ is negligible, that is, when $\tau_k$ is small, for instance, when $K$ is large, then $d$ will be large and $t_d$ will be close to the standard normal distribution, implying that our proposed $t$-test will reduce to the usual $z$-test.

### 3.2. Simulation

Simulation studies were conducted to investigate the performance of the proposed $t$-test. We consider a random-effects logistic model:

$$\mathrm{logit}(\mu_{ij}|b_i)=x_{ij}\beta_1+b_i$$

where $x_{ij}$ are i.i.d. from a Bernoulli distribution $\mathrm{Bin}(1,1/2)$, and $b_i$ are i.i.d. from a standard normal distribution $\mathrm{N}(0,1)$, and they are independent of each other. Conditional on $b_i$, $y_{ij}$'s have independent Bernoulli distributions $\mathrm{Bin}(1,\mu_{ij})$ with $j=1,2,\ldots,20$ and $i=1,\ldots,K$ ($K=$ 10, 20 or 30). Note that, in general, a non-linear random-effects model 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 [11], and it is much easier to generate simulated data according to a random-effects model. For each case (determined by $K$ and $\beta$ described below), 500 samples were generated, and the correct marginal logistic regression model was fitted with the use of the working independence model in GEE.

First we consider the size properties of the $z$- and $t$-tests when $\beta_1=0$. The empirical size/power based on 500 simulations, and 95 per cent normal confidence interval of the size/power for each set-up are presented in Table I. Note that we always truncate the lower endpoint of a confidence interval at 0 if it is smaller than 0. It is verified that when $K$ is as small as 10, the size of the normal-based test is much larger than the specified nominal levels 0.05 or 0.01. As $K$ increases, the size of the normal-based test gets closer and closer

 

Table I. Empirical size and power (and their 95 per cent confidence intervals) of the $\alpha$-level $z$- and $t$-tests for testing $H_0 : \beta_1 = 0$ in a mixed-effects logistic model.

| Set-up | | $z$-test | | $t$-test | |
|---|---|---|---|---|---|
| $\beta_1$ | $K$ | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ |
| 0 | 10 | 0.090 (0.065, 0.115) | 0.034 (0.018, 0.050) | 0.064 (0.042, 0.085) | 0.016 (0.005, 0.027) |
| 0 | 20 | 0.054 (0.034, 0.074) | 0.024 (0.011, 0.037) | 0.048 (0.029, 0.067) | 0.012 (0.002, 0.022) |
| 0 | 30 | 0.056 (0.036, 0.076) | 0.008 (0.000, 0.016) | 0.050 (0.031, 0.069) | 0.006 (0.000, 0.013) |
| 0.4 | 30 | 0.512 (0.468, 0.556) | 0.284 (0.244, 0.324) | 0.494 (0.450, 0.538) | 0.252 (0.214, 0.290) |
| 0.5 | 30 | 0.704 (0.664, 0.744) | 0.480 (0.436, 0.524) | 0.666 (0.625, 0.707) | 0.414 (0.371, 0.457) |
| 0.6 | 30 | 0.844 (0.812, 0.876) | 0.650 (0.608, 0.692) | 0.828 (0.795, 0.861) | 0.592 (0.549, 0.635) |
| 0.7 | 30 | 0.942 (0.922, 0.962) | 0.818 (0.784, 0.852) | 0.928 (0.905, 0.951) | 0.782 (0.746, 0.818) |

to the nominal levels (due to the consistency of the sandwich variance estimator). It can be also seen clearly that in many situations the size of the $t$-based test is closer to the nominal levels than that of the normal-based test.

When $K = 30$, both tests have satisfactory size. Then it is interesting to investigate whether there is a significant loss of power using the $t$-test. With various values of $\beta_1$, we see that the power difference of the two tests is often small; the largest difference is about 7 per cent (Table I). In summary, compared with the $z$-test, our proposed $t$-test has size close to the nominal levels while maintaining reasonable power.

### 3.3. Remarks

In the above simulations, we considered the normal random effects $b_i$. We also did some simulations with $b_i$ generated from a $t_3$ distribution, and similar results (not shown) were obtained as before.

In Section 2 we propose estimating the variability of the sandwich estimator by ignoring the variability of the model-based variance estimator. The simulation results confirm that this is a sensible approach. In the simulations, we found that the ratio of the (sample) variances of the model-based variance estimates and the sandwich estimates is only around 1 per cent, implying that the variability of the model-based variance estimator is relatively negligible. On the other hand, a straightforward way to estimate the variability of the whole sandwich estimator is to use the jack-knife (see reference [12]). Although using the jack-knife is computationally more intensive, it does not seem to have better performance than the current proposal in a simulation study. For instance, as a comparison with those in Table I, for the set-ups $K = 10$ and $K = 30$ as in Section 3.2, the size of the 5 per cent (1 per cent) level $t$-test based on

 

the jack-knife method is 0.034 (0.002) and 0.048 (0.004), respectively; for the set-up $K=30$ and $\beta_1=0.5$, the power of the new $t$-test is 0.652 (0.372), slightly smaller than that of the original $t$-test.

An asymptotically equivalent test to the Wald test is the generalized score test [13, 14]. However, the Wald method is also useful in providing confidence intervals. In fact, at this moment, only the Wald test is implemented in many statistical packages, such as SAS and S-plus. It is possible to extend our idea of adjusting for the variability of the sandwich estimator to the generalized score test (but we do not pursue it here). Owing to the correspondence between the Wald test and the Wald confidence interval, we expect that the Wald confidence interval using our proposed $t$ coefficient has a higher and closer to the nominal level coverage percentage than does that based on the normal coefficient.

## 4. APPROXIMATE $F$-TEST

### 4.1. Derivation

Without loss of generality (see also Section 4.3), we consider testing $H_0:\beta=0$. We propose to approximate the distribution of $vV_S$ by a Wishart distribution $\mathcal{W}_p(v,\Sigma)$ with degrees of freedom $v$, dispersion matrix $\Sigma$ and dimension $p\times p$ (for example, Vonesh and Chinchilli, reference [15], pp. 25–26). If we assume that $vV_S\sim\mathcal{W}_p(v,\Sigma)$, we have (for example, Vonesh and Chinchilli, reference [15], p. 26)

$$E(vV_S)=v\Sigma,\quad \operatorname{cov}(\operatorname{vec}(vV_S))=v(I_{p^2}+I_{(p,p)})(\Sigma\otimes\Sigma)$$

where $I_{p^2}$ is a $p^2\times p^2$ identity matrix and

$$I_{(p,p)}=\begin{pmatrix} E'_{11} & E'_{12} & \dots & E'_{1p}\\ E'_{21} & E'_{22} & \dots & E'_{2p}\\ \dots & & & \\ E'_{p1} & E'_{p2} & \dots & E'_{pp}\end{pmatrix}$$

with $E_{jk}$ being a $p\times p$ matrix of zeros except that its $(j,k)$th element is 1. Hence under $vV_S\sim\mathcal{W}_p(v,\Sigma)$, $V_S$ is a reasonable estimator for $\Sigma$, and an estimator for $\operatorname{cov}(\operatorname{vec}(vV_S))$ is

$$\hat{\Omega}=v(I_{p^2}+I_{(p,p)})(V_S\otimes V_S) \tag{4}$$

The degrees of freedom $v$ can be chosen such that the empirical covariance matrix of $vV_S$

$$\widehat{\operatorname{cov}}(\operatorname{vec}(vV_S))=v^2\widehat{\operatorname{cov}}(\operatorname{vec}(V_S))$$

with $\widehat{\operatorname{cov}}(\operatorname{vec}(V_S))$ given in (3), is close to $\hat{\Omega}$, the estimated covariance matrix of $vV_S$ under the Wishart approximation. Thus in our implementation, $v$ minimizes the sum of squared errors between $v\operatorname{vec}(\widehat{\operatorname{cov}}(\operatorname{vec}(V_S)))$ and $\operatorname{vec}(\hat{\Omega})/v$. As a referee pointed out, an alternative way to find $v$ is to minimize the sum of squared errors between $v\operatorname{vech}(\widehat{\operatorname{cov}}(\operatorname{vech}(V_S)))$ and $\operatorname{vech}(\hat{\Omega}^*)/v$, where $\hat{\Omega}^*$ estimates $\operatorname{cov}(\operatorname{vech}(V_S))$ from the Wishart model and $\operatorname{vech}(V_S)$ only takes upper submatrix of $V_S$ (see Vonesh and Chinchilli, reference [15], p. 12). This alternative and other more robust regression methods may be more effective and warrant future studies.

 

Since Wald's chi-squared statistic is $W = v[\hat{\beta}'(vV_S)^{-1}\hat{\beta}]$, where $\hat{\beta}$ is approximately distributed as $N_p(0, \Sigma)$ under $H_0$, if $\hat{\beta}$ and $vV_S$ are approximately independent, then $W$ approximately has the same distribution as Hotelling's $T^2$ (for example, Vonesh and Chinchilli, reference [15], p. 30), and therefore

$$\frac{v - p + 1}{vp} W \sim F(p, v - p + 1)$$

under $H_0$. Rather than using $\chi^2_p$, we propose using the above scaled $F$-distribution as the reference for $W$ to calculate $p$-values.

Note that if the variability of $V_S$ is very small, it can be verified that $v$ will be very large, then our proposed $F$-test reduces to the usual Wald chi-squared test. In addition, if $p = 1$ (that is, testing a single parameter), it is easy to verify that the above $F$-test is equivalent to the $t$-test discussed earlier. In the following, the degrees of freedom parameter $v$ of the $F$-test is derived based on $p > 1$, and it can be also used to test each individual parameter in $H_0$, but then note that the resulting $F$-test is not equivalent to the $t$-test in Section 3 since their degrees of freedom are calculated in different ways.

### 4.2. Simulation

Simulated data were generated from a model similar to that used in Section 3.2 except that there are three covariates:

$$\text{logit}(\mu_{ij}|b_i) = x_{ij1}\beta_1 + x_{ij2}\beta_2 + x_{ij3}\beta_3 + b_i$$

where the covariates $x_{ij1}$, $x_{ij2}$ and $x_{ij13}$ are generated as i.i.d. from a Bernoulli distribution Bin$(1, 1/2)$, $b_i$'s are i.i.d. from N$(0, 1)$, and they are independent of each other. We again take $n = 20$ and consider four cases where $K = 10$, 20, 30 and 40. The true values of the regression coefficients are $\beta_1 = \beta_2 = \beta_3 = 0$, but the above model with three covariates is always fitted.

The null hypothesis to be tested is either $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ or $H_0 : \beta_1 = 0$. Note that for the latter the same degrees of freedom $v$ is used as that for the former. The rejection proportions from Wald's chi-squared test and our proposed $F$-test are presented in Table II. It is obvious that the chi-squared test has dramatically inflated type I errors, whereas the $F$-test has well-controlled type I errors and sometimes appears to be conservative. In general, the performance of the Wald chi-squared test is worse in testing multiple parameters than that in testing a single parameter. Note that even for $K$ as large as 40, the Wald chi-squared test may still have type I errors much larger than the nominal levels.

As explained at the end of Section 4.1, in testing $H_0 : \beta_1 = 0$, our proposed $F$-test and $t$-test may not be equivalent: the denominator degrees of freedom in the $F$-test is calculated based on the submatrix of the sandwich estimate for $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)'$, whereas in the $t$-test it only uses the sandwich variance estimate for $\hat{\beta}_1$. Although the $F$-test is more general, in testing a single parameter there is no apparent advantage of using the $F$-test over using the $t$-test. Comparing Tables I and II, it seems that the $F$-test is more conservative than the $t$-test.

Finally we note that our proposed $F$-test can be readily extended to more general situations with a general null hypothesis $H_0 : L\beta = 0$, where $L$ is an $r \times p$ matrix with full rank $r \leqslant p$. The Wald statistic is $W = (L\hat{\beta})'[LV_SL']^{-1}(L\hat{\beta})$. As in Section 4.1, we can approximate $LV_SL'$ by a scaled Wishart distribution, and hence approximate $W$ by a scaled $F$-distribution. In particular,

 

Table II. Empirical size (and its 95 per cent confidence interval) of the $\alpha$-level Wald $\chi^2$ and proposed $F$-tests for testing $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ or $H_0: \beta_1 = 0$ in a mixed-effects logistic model.

| Set-up | | $\chi^2$ test | | $F$-test | |
|---|---|---|---|---|---|
| $H_0$ | $K$ | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ |
| $\beta_1 = \beta_2 = \beta_3 = 0$ | 10 | 0.240 (0.203, 0.277) | 0.136 (0.106, 0.166) | 0.034 (0.018, 0.050) | 0.006 (0.000, 0.013) |
| | 20 | 0.120 (0.092, 0.148) | 0.050 (0.031, 0.069) | 0.026 (0.012, 0.040) | 0.002 (0.000, 0.006) |
| | 30 | 0.096 (0.070, 0.122) | 0.026 (0.012, 0.040) | 0.026 (0.012, 0.040) | 0.004 (0.000, 0.010) |
| | 40 | 0.084 (0.060, 0.108) | 0.024 (0.011, 0.037) | 0.036 (0.020, 0.052) | 0.006 (0.000, 0.013) |
| $\beta_1 = 0$ | 10 | 0.100 (0.074, 0.126) | 0.032 (0.017, 0.047) | 0.036 (0.020, 0.052) | 0.008 (0.000, 0.016) |
| | 20 | 0.066 (0.044, 0.088) | 0.010 (0.001, 0.019) | 0.032 (0.017, 0.047) | 0.002 (0.000, 0.006) |
| | 30 | 0.062 (0.041, 0.083) | 0.016 (0.005, 0.027) | 0.044 (0.026, 0.062) | 0.006 (0.000, 0.013) |
| | 40 | 0.068 (0.046, 0.090) | 0.026 (0.012, 0.040) | 0.056 (0.036, 0.076) | 0.014 (0.004, 0.024) |

if we are testing a contrast of regression coefficients (that is, $r = 1$), an approximate $t$-test can be also constructed as in Section 3.1 where $LV_S L'$ can then be approximated by a scaled chi-squared distribution.

### 4.3. A comparison with other approaches

Awareness of the unsatisfactory small-sample performance of the usual Wald chi-squared test using the sandwich estimator has arisen and several approaches have been proposed very recently. Fay *et al.* [6] pointed out two general ways in a different setting: one is to take account of the variability of the sandwich estimator, and the second is to correct the bias of the sandwich estimator. We feel that to some extent bias correction is helpful (see also reference [7]), but a more effective way is to account for the variability of the sandwich estimator. This is the approach we have taken and its advantage will be shown in our simulation studies. Pan [8] also proposed a modification to the sandwich estimator based on slightly stronger conditions. Here we briefly introduce the alternative approaches and then compare them with our proposal.

Mancl and DeRouen [7] proposed to correct the bias of the sandwich estimator $V_S$ in (2) by replacing $S_i S_i'$ with

$$(I - H_i)^{-1} S_i S_i' (I - H_i')^{-1}$$

 

where $I$ is an $n \times n$ identity matrix, and $H_i = D_i V_M D_i' V_i^{-1}$. The motivation is that $S_i S_i'$ is a biased estimator of $\mathrm{cov}(Y_i)$; that is, $E(S_i S_i') \approx (I - H_i)\mathrm{cov}(Y_i)(I - H_i')$. We denote this bias-corrected sandwich estimator as $V_{BC}$.

It is natural to combine our proposal with that of Mancl and DeRouen [7] to account for both the variability and bias of the sandwich estimator. This can be easily implemented; everything is the same as before except replacing $S_i$ of $V_S$ in our proposal by its bias-corrected version $(I - H_i)^{-1} S_i$ of $V_{BC}$. However, since our original $F$-test (before doing bias correction) is slightly conservative, it may not be of much use to further correct the downward bias of the sandwich estimator. This point will be verified next.

Pan [8] observed that it is effective to modify the sandwich estimator by estimating a common correlation matrix $\mathrm{corr}(Y_i)$ if it exists (or more generally, if there is a common correlation matrix for each of a small number of subject groups). Note that usually $\mathrm{var}(Y_{ij})$ can be modelled well as in GLMs for independent data. Then the common correlation matrix $\mathrm{corr}(Y_i)$ can be estimated using the unstructured correlation matrix estimator

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

Thus $\mathrm{cov}(Y_i)$ can be estimated by

$$W_i = \phi A_i^{1/2} R_U A_i^{1/2} = A_i^{1/2} \left( \sum_{i=1}^{K} A_i^{-1/2} S_i S_i' A_i^{-1/2} / K \right) A_i^{1/2} \tag{5}$$

Note that $W_i$ does not depend on $\phi$, and it may be a better estimator of $\mathrm{cov}(Y_i)$ than is $S_i S_i'$. Replacing $S_i S_i'$ in (2) by $W_i$, Pan [8] obtained a new sandwich estimator, denoted by $V_N$.

We can conduct the Wald chi-squared test using either $V_{BC}$ or $V_N$, rather than the usual $V_S$. We can also use the $F$-test based on $V_{BC}$. Following the same simulation set-up as in Section 4.2, these different methods are applied and the results are listed in Table III. Note that one empirical size in Table III is 0, and we give Louis' [16] one-sided confidence interval.

Comparing Tables II and III, we can verify that bias correction improves the performance of the Wald chi-squared test. However, for $K$ as small as 10 or 20, the Wald test based on $V_{BC}$ may still have much inflated type I errors, especially in testing for more than one parameter. Hence, for small $K$, our proposed $F$-test is advantageous in maintaining proper size. On the other hand, the $F$-test that accounts for both the variability and bias of the sandwich estimator appears too conservative, which can be seen by the type I errors of $F$-$V_{BC}$ often being much smaller than the nominal levels. This is not surprising since the original $F$-test (Table II) is already slightly conservative. Hence, it does not seem necessary to do bias correction in our proposed $F$-test.

In general, the performance of using $V_N$ is satisfactory. However, it is reminded that $V_N$ is based on stronger assumptions, and in some situations, such as when the time points of observing subjects are largely different, the assumption of having a common correlation structure may be in question. Thus our proposed $F$-test is attractive for its performance as well as its flexibility.

 

Table III. Empirical size (and its 95 per cent confidence interval) of the α-level Wald $\chi^2$ tests based on using modified sandwich estimators $V_{BC}$ and $V_N$, and the $F$-test based on $V_{BC}$, for testing $H_0: \beta_1=\beta_2=\beta_3=0$ or $H_0: \beta_1=0$ in a mixed-effects logistic model.

| Set-up | | $\chi^2$-$V_{BC}$ | | $F$-$V_{BC}$ | | $\chi^2$-$V_N$ | |
|---|---|---|---|---|---|---|---|
| $H_0$ | $K$ | $\alpha=0.05$ | $\alpha=0.01$ | $\alpha=0.05$ | $\alpha=0.01$ | $\alpha=0.05$ | $\alpha=0.01$ |
| $\beta_1=\beta_2=\beta_3=0$ | 10 | 0.160 (0.128, 0.192) | 0.088 (0.063, 0.113) | 0.016 (0.005, 0.027) | 0.004 (0.000, 0.010) | 0.066 (0.044, 0.088) | 0.012 (0.002, 0.022) |
| | 20 | 0.096 (0.070, 0.122) | 0.028 (0.014, 0.042) | 0.016 (0.005, 0.027) | 0.002 (0.000, 0.006) | 0.056 (0.036, 0.076) | 0.002 (0.000, 0.006) |
| | 30 | 0.066 (0.044, 0.088) | 0.014 (0.004, 0.024) | 0.016 (0.005, 0.027) | 0.000 (0.000, 0.006) | 0.046 (0.028, 0.064) | 0.004 (0.000, 0.010) |
| | 40 | 0.068 (0.046,0.090) | 0.018 (0.006, 0.030) | 0.026 (0.012, 0.040) | 0.002 (0.000, 0.006) | 0.058 (0.038, 0.078) | 0.010 (0.001, 0.019) |
| $\beta_1=0$ | 10 | 0.060 (0.039, 0.081) | 0.022 (0.009, 0.035) | 0.022 (0.009, 0.035) | 0.004 (0.000, 0.010) | 0.054 (0.034, 0.074) | 0.016 (0.005, 0.027) |
| | 20 | 0.048 (0.029, 0.067) | 0.008 (0.000, 0.016) | 0.026 (0.012, 0.040) | 0.002 (0.000, 0.006) | 0.044 (0.026, 0.062) | 0.002 (0.000, 0.006) |
| | 30 | 0.050 (0.031, 0.069) | 0.010 (0.001, 0.019) | 0.034 (0.018, 0.050) | 0.004 (0.000, 0.010) | 0.030 (0.015, 0.045) | 0.008 (0.000, 0.016) |
| | 40 | 0.062 (0.041, 0.083) | 0.020 (0.008, 0.032) | 0.048 (0.029, 0.067) | 0.010 (0.001, 0.019) | 0.064 (0.043, 0.085) | 0.018 (0.006, 0.030) |

## 5. EXAMPLES

### 5.1. A 4×4 cross-over trial with a continuous response

Fleiss (reference [17], Table 10.12) listed and analysed a data set from a $4 \times 4$ cross-over trial comparing three active treatments with a control. There were total 20 subjects, each receiving each treatment for 1 week. Williams' restricted Latin square design was used, where each treatment follows each of the others the same number of times (for example, Cochran and Cox, reference [18], pp. 133–139). The response variable (that is, dental plaque score) is continuous. Let $Y_{ijkl}$ denote the response value in period $i$ on treatment $j$ when treatment $k$ was given in the preceding period for subject $l$, where $i=1,\ldots,4$, $j=1,\ldots,4$, $k=1,\ldots,4$ and $l=1,\ldots,20$. Let $\mu$ denote the intercept term, $\pi_i$ the effect of the $i$th period, $\tau_j$ the $j$th treatment effect, $\alpha_k$ the residual effect from treatment $k$, $b_l$ the $l$th subject's random effect, and $\varepsilon_{ijkl}$ the random error. Then the model is

$$Y_{ijkl} = \mu + \pi_i + \tau_j + \alpha_k + b_l + \varepsilon_{ijkl}$$

where $\pi_4 = \tau_4 = \alpha_4 = 0$, $b_l$'s are i.i.d. from $N(0, \sigma_P^2)$, $\varepsilon_{ijkl}$'s are i.i.d. from $N(0, \sigma^2)$, and $b_l$ and $\varepsilon_{ijkl}$ are independent.

To test whether there is treatment effect (that is, $H_0 : \tau_1 = \tau_2 = \tau_3 = 0$), the ANOVA $F$-test yields the $p$-value 0.2845 (Fleiss, reference [17], p. 286). Now we consider fitting a linear marginal model using GEE. The usual Wald chi-square test and our $F$-test result in the $p$-values of 0.1077 and 0.3198, respectively. The estimated denominator degrees of freedom of our $F$-test is 9.4. Apparently our $F$-test leads to the result closer to that based on the traditional parametric method. Furthermore, based on the fitted model for $\pi_i$'s, $\alpha_i$'s and variance components, but using $\tau_1 = \tau_2 = \tau_3 = 0$, we did 1000 simulations, and found that the type I errors of the 5 per cent or 1 per cent Wald chi-square test are, respectively, 14.2 per cent and 5.0 per cent. Hence, we conclude that for this data set, the Wald chi-square test may be too liberal.

### 5.2. The Lung Health Study with a binary response

We consider the data from the Lung Health Study (LHS) [19] for illustration. The LHS was a multi-centre randomized controlled clinical trial. The participants were all smokers between the ages of 35 and 60 years 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). A behavioural intervention programme was provided to all participants in the two intervention groups to encourage and help them quit smoking. The participants were followed for five years. Our goal here is to investigate whether there is any intervention effect in reducing the smoking rate. The response variable $y_{it}$ is quitting status of participant $i$ at year $t$, which is 1 for a quitter and 0 otherwise. We consider a marginal logistic regression model

$$\text{logit}(\pi_{it}) = \beta_0 + \beta_1 I(t=1) + \beta_2 \text{SIA}_i + \beta_3 \text{SIP}_i \tag{6}$$

where $\text{SIA}_i$ or $\text{SIP}_i$ is an indicator of whether participant $i$ is in group SIA or SIP, respectively ($=1$ if yes; $=0$ otherwise), $\pi_{it} = E(y_{it}|\text{SIA}_i, \text{SIP}_i)$ for $t=1,2,\ldots,5$ and $i=1,2,\ldots,K$. It was found that the quitting rate at year 1 is significantly different the from other four years, hence

 

Table IV. $P$-values of the various tests for the LHS data.

| $K$ | $H_0:\beta_2=0$ | | | $H_0:\beta_3=0$ | | | $H_0:\beta_2=\beta_3=0$ | |
|---|---|---|---|---|---|---|---|---|
| | $\chi^2$ | $t$ | $F$ | $\chi^2$ | $t$ | $F$ | $\chi^2$ | $F$ |
| 15 | 0.059 | 0.086 | 0.130 | 0.153 | 0.186 | 0.224 | 0.160 | 0.370 |
| 30 | 0.010 | 0.028 | 0.055 | 0.471 | 0.485 | 0.506 | 0.025 | 0.183 |
| 60 | 0.002 | 0.004 | 0.007 | 0.189 | 0.196 | 0.205 | 0.009 | 0.028 |
| 90 | 0.002 | 0.003 | 0.005 | 0.030 | 0.035 | 0.043 | 0.006 | 0.021 |
| 120 | 0.000 | 0.000 | 0.003 | 0.036 | 0.040 | 0.047 | 0.004 | 0.015 |

a binary covariate indicating year 1 is also included. Taking several random samples with various sizes $K$, each of which consists of an equal number of participants in each treatment group, we obtain the results shown in Table IV to test three null hypotheses: (i) $H_0:\beta_2=0$ (that is, no difference between SIA and UC); (ii) $H_0:\beta_3=0$ (that is, no difference between SIP and UC); (iii) $H_0:\beta_2=\beta_3=0$ (that is, no difference among the three groups). For smaller sample sizes $K$, it is less likely to draw significant conclusions from the approximate $t$- or $F$-test than it is from the Wald chi-squared test.

## 6. DISCUSSION

Our numerical studies have confirmed that using the sandwich (co)variance estimator in the Wald chi-squared test can lead to dramatically inflated type I errors when the sample size is not large. The reason is due to the well-known fact that the sandwich variance estimator has large variation with small samples. The non-negligible variability of the sandwich estimator is not taken into account in the Wald chi-squared test. In this article, we proposed using approximate $t$- and $F$-tests with adjustment for the variability of the sandwich estimator in the Wald statistic. We found that the proposed tests obtain type I errors closer to the nominal levels than does the usual chi-squared test. Note that the idea of accounting for the variability of variance estimates is not new and has a long history (for example, reference [9]), though to our knowledge it has never been discussed in the context of GEE. The adjustment has proved to be useful in many other applications, particularly in the linear mixed-effects models [20, 21]. However, there are two interesting features in our proposal. First, due to the empirical nature of the sandwich covariance estimator, we propose accordingly to use the sample covariance matrix to estimate its variability. Second, in the approximate $F$-test, its denominator degrees of freedom is determined by matching the first two moments of the sandwich covariance estimator with those of a scaled Wishart variate. This latter point is not considered by Fay *et al.* [6]. Interestingly, the resulting test statistic is the same as the (scaled) Wald statistic, though approximate $t$- or $F$-distribution, rather than a chi-squared distribution, is used as the reference to calculate $p$-values. Owing to the equivalence of a Wald test and its corresponding confidence interval (or region), our result implies that using our proposed $t$- or $F$-coefficient will improve the coverage percentage of the resulting confidence interval (or region). Finally, in terms of the trade-off between the power and type I error, our general view is that caution should be exercised when using tests with possibly inflated type I errors. Hence we prefer our proposed $t$- or $F$-test to the usual chi-squared test since the former often has type I errors

 

closer to the nominal level than the latter for small samples, and they will agree with each other as the sample size increases. It would be of interest to investigate in the future how to further improve the performance of the $t$- or $F$-test.

ACKNOWLEDGEMENTS

We are grateful to two referees for many helpful and constructive comments. We thank Dr John Connett for many helpful discussions on the LHS and relevant issues.

REFERENCES

1. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
2. Prentice RL. Correlated binary regression with covariates specific to each binary observation. *Biometrics* 1988; **44**:1033–1048.
3. Drum M, McCullagh P. Comment. *Statistical Science* 1993; **8**:300–301.
4. Emrich LJ, Piedmonte MR. On some small sample properties of generalized estimating equation estimates for multivariate dichotomous outcomes. *Journal of Statistical Computation and Simulation* 1992; **41**:19–29.
5. Gunsolley JC, Getchell C, Chinchilli VM. Small sample characteristics of generalized estimating equations. *Communications in Statistics–Simulation* 1995; **24**:869–878.
6. Fay MP, Graubard BI, Freedman LS, Midthune DN. Conditional logistic regression with sandwich estimators: application to a meta analysis. *Biometrics* 1998; **54**:195–208.
7. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; **57**:126–134.
8. Pan W. On the robust variance estimator in generalised estimating equations. *Biometrika* 2001; **88**:901–906.
9. Satterthwaite FF. Synthesis of variance. *Psychometrika* 1941; **6**:309–316.
10. Cox DR, Hinkley DV. *Theoretical Statistics*. Chapman and Hall: London, 1974.
11. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**:1049–1060.
12. Shao J, Tu DS. *The Jackknife and Bootstrap*. Springer: New York, 1995.
13. Rotnitzky A, Jewell NP. Hypothesis testing of regression parameters in semiparametric generalized linear models for clustered data. *Biometrika* 1990; **77**:485–497.
14. Lefkopoulou M, Ryan L. Global tests for multiple binary outcomes. *Biometrics* 1993; **49**:975–988.
15. Vonesh EF, Chinchilli VM. *Linear and Nonlinear Models for the Analysis of Repeated Measurements.* Marcel Dekker: New York, 1997.
16. Louis TA. Confidence intervals for a binomial parameter after observing no success. *American Statistician* 1981; **35**:154.
17. Fleiss J. *The Design and Analysis of Clinical Experiments*. Wiley: New York, 1986.
18. Cochran WG, Cox GM. *Experimental Designs*. 2nd edn. Wiley: New York, 1957.
19. Connett JE, Kusek JW, Bailey WC, O'Hara P, Wu M for the Lung Health Study Research Group. Design of the Lung Health Study: a randomized clinical trial of early intervention for chronic obstructive pulmonary disease. *Controlled Clinical Trials* 1993; **14**:3S–19S.
20. Kackar RN, Harville DA. Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association* 1984; **79**:853–862.
21. Giesbrecht FG, Burns JC. Two-stage analysis based on a mixed model: large-sample asymptotic theory and small-sample simulation results. *Biometrics* 1985; **41**:477–486.

 