Goodness-of-fit Tests for GEE with Correlated Binary Data

WEI PAN

University of Minnesota

ABSTRACT. The marginal logistic regression, in combination with GEE, is an increasingly important method in dealing with correlated binary data. As for independent binary data, when the number of possible combinations of the covariate values in a logistic regression model is much larger than the sample size, such as when the logistic model contains at least one continuous covariate, many existing chi-square goodness-of-fit tests either are not applicable or have some serious drawbacks. In this paper two residual based normal goodness-of-fit test statistics are proposed: the Pearson chi-square and an unweighted sum of residual squares. Easy-to-calculate approximations to the mean and variance of either statistic are also given. Their performance, in terms of both size and power, was satisfactory in our simulation studies. For illustration we apply them to a real data set.

Key words: generalized estimating equations, logistic regression, Pearson's chi-square, unweighted sum of squares

1. Introduction

The logistic regression model is widely used in analyzing independent binary response data. In familial or longitudinal studies, correlated binary data often arise. The marginal logistic regression and its associated generalized estimating equations (GEE) method are an increasingly important approach to analysing such data (Liang & Zeger, 1986). In general, it is important to assess the overall goodness-of-fit of the regression model. Recently, two methods were proposed for correlated binary data (Barnhart & Williamson, 1998; Horton *et al*., 1999). These two tests can be regarded respectively as an extension of the goodness-of-fit test of Tsiatis (1980) and that of Hosmer & Lemeshow (1980) for ordinary logistic regression (with independent binary data) to marginal logistic regression (with correlated binary data). They are based on forming groups by partitioning the space of covariates or of predicted probabilities such that statistical inference can be drawn based on the chi-square distribution. We believe that these tests are useful in practice. Their disadvantage is that the resulting test statistics depend on the subjective partitioning. Hosmer *et al*. (1997) gave an example for ordinary logistic regression. Using a well known-data set, they showed that six commonly used statistical packages give six different p-values for the Hosmer–Lemeshow test: four packages produce a statistic with a p-value > 0.1, one with $0.05 < p$-value < 0.1 and one with a p-value < 0.05. This unpleasant discrepancy among them arises from their difference in forming groups to construct the statistic. Presumably this issue will also persist with a test based on partitioning in the context of marginal modelling.

In this paper, we consider the situation where the number of possible combinations of the covariate values in a logistic model is much larger than the sample size, such as when the logistic model contains at least one continuous covariate. Our approach is a natural extension of that for independent binary data to correlated binary data. In a nice review of many existing goodness-of-fit tests for ordinary logistic regression, Hosmer *et al*. (1997) demonstrated that the asymptotic normal Pearson's chi-square statistic and an unweighted sum of residual squares statistic enjoy good properties of both power and size. Note that for binary data without

grouping, the Pearson chi-square does not have a usual chi-square distribution. It seems desirable to extend these two test statistics to marginal logistic regression. We pursue it here.

2. Two goodness-of-fit tests

Suppose we have N subjects (or clusters). For each subject $i, 1 \le i \le N$, there are N_i binary (0 or 1) response values $y_i = (y_{i1}, \ldots, y_{iN_i})'$ and covariate matrix $x_i = (x_{i1}, \ldots, x_{iN_i})'$. From now on, for simplicity we assume $N_i = m$ for all i. For $i \neq j$, y_i and y_j are independent, but generally the components of each y_i are correlated. The marginal logistic regression model specifies that $\text{logit}(\pi_{it}) = x_{it}'\beta$, where $\pi_{it} = E(y_{it}|x_{it})$ and $\text{var}(y_{it}|x_{it}) = \pi_{it}(1 - \pi_{it})$. The unknown regression coefficient (vector) β is of primary interest, which can be estimated through solving the following generalized estimating equations (Liang & Zeger 1986):

$$S(\beta) = \sum_{i=1}^{N} \left(\frac{\partial \pi_i}{\partial \beta'}\right)' V_i^{-1}(y_i - \pi_i) = \sum_{i=1}^{N} x_i' A_i V_i^{-1}(y_i - \pi_i) = 0, \tag{1}$$

where A_i is a diagonal matrix $\text{diag}[\pi_{i1}(1 - \pi_{i1}), \ldots, \pi_{im}(1 - \pi_{im})]$, and V_i is the working covariance matrix of y_i. V_i can be expressed in terms of the working correlation matrix $R_W = R_W(\alpha)$: $V_i = A_i^{1/2} R_W A_i^{1/2}$, where α may be some unknown parameters involved in the working correlation structure, which can be estimated through moment methods or another set of estimating equations. An attractive point of GEE is that it can yield consistent and asymptotically normal estimate of β, $\hat{\beta} = \hat{\beta}(R_W)$, even when the working correlation matrix R_W is incorrectly specified. For instance, we can use the working independence model with $R_W = I$, the identity matrix. The choice of R_W will influence the estimation efficiency: it is more efficient to use R_W that is closer to the true underlying correlation structure. Many studies have shown that $\hat{\beta}$ obtained under the working independence model is relatively efficient (Zeger, 1988; McDonald, 1993; Sutradhar & Das, 1999), at least when the within-subject correlation is not too strong.

For any expression $B = B(\beta)$, we write $\hat{B} = B(\hat{\beta})$. For instance, $\hat{\pi}_{ij} = \text{logit}^{-1}(x_{ij}'\hat{\beta})$ is the estimate of π_{ij} by replacing β with $\hat{\beta}$. We also use the conventional notation that $Y = (y_1', \ldots, y_N')'$, $X = (x_i, \ldots, x_N)'$, $A = \text{diag}(A_1, \ldots, A_N)$, $V = \text{diag}(V_1, \ldots, V_N)$, and similarly for π and others.

Now we derive an approximation to residuals $\hat{e} = Y - \hat{\pi}$. A Taylor expansion of $\hat{e}$ at the true β^* leads to

$$Y - \hat{\pi} \approx Y - \pi - \left[\left.\frac{\partial \pi}{\partial \beta'}\right|_{\beta=\beta^*}\right](\hat{\beta} - \beta^*) = Y - \pi - AX(\hat{\beta} - \beta^*). \tag{2}$$

By another Taylor expansion of $S(\hat{\beta})$ at β^* we have

$$\begin{aligned} 0 = S(\hat{\beta}) &\approx S(\beta^*) + \left[\left.\frac{\partial S}{\partial \beta'}\right|_{\beta=\beta^*}\right](\hat{\beta} - \beta^*) \\ &= \sum_{i=1}^{N} x_i' A_i V_i^{-1}(y_i - \pi_i) + \left[\left.\frac{\partial S}{\partial \beta'}\right|_{\beta=\beta^*}\right](\hat{\beta} - \beta^*), \end{aligned} \tag{3}$$

where a general (but complex) formula for $\partial S/\partial \beta'$ is given in the appendix. If we approximate $A_i V_i^{-1}$ as a constant matrix, which is true if we use the working independence model $V = A$, then $\partial S/\partial \beta'$ can be largely simplified as

$$\frac{\partial S}{\partial \beta'} = -\sum_{i=1}^{N} x_i' A_i V_i^{-1} A_i x_i. \tag{4}$$

Combining (2)–(4), we have

$$Y - \hat{\pi} \approx (I - H)(Y - \pi), \quad \text{where } H = AX(X'AV^{-1}AX)^{-1}X'AV^{-1}. \tag{5}$$

If we use the working independence model $V = A$, H reduces to $H = AX(X'AX)^{-1}X'$, which is the same as that given in Hosmer *et al.* (1997) for independent binary data.

The Pearson chi-square statistic is

$$\begin{aligned} G &= \sum_{i=1}^{N}\sum_{j=1}^{m}\frac{(y_{ij} - \hat{\pi}_{ij})^2}{\hat{\pi}_{ij}(1 - \hat{\pi}_{ij})} \\ &= Nm + (1 - 2\hat{\pi})'\hat{A}^{-1}\hat{e} \approx Nm + (1 - 2\hat{\pi})'\hat{A}^{-1}(I - H)e. \end{aligned}$$

Treating $(1 - 2\hat{\pi})'\hat{A}^{-1}$ as fixed, we obtain the approximate mean and variance of G:

$$\widehat{E(G)} = Nm, \quad \widehat{\text{var}(G)} = (1 - 2\hat{\pi})'\hat{A}^{-1}(I - \hat{H})\widehat{\text{cov}(Y)}(I - \hat{H}')\hat{A}^{-1}(1 - 2\hat{\pi}).$$

There are two ways to estimate $\text{cov}(Y)$. The first is to use the empirical covariance estimator $\widehat{\text{cov}(Y)}_e = \text{diag}(\widehat{\text{cov}(Y_1)}, \ldots, \widehat{\text{cov}(Y_N)})$, where $\widehat{\text{cov}(Y_i)} = (y_i - \hat{\pi}_i)(y_i - \hat{\pi}_i)'$, as used in the robust covariance estimator of the estimated regression coefficients $\hat{\beta}$ (Liang & Zeger, 1986). Its advantage is the simplicity and generality. However, as a crude estimator it may not be efficient. Hence we propose to use the second, $\widehat{\text{cov}(Y)}_u = \hat{A}^{1/2}\text{diag}(\hat{R}_u, \ldots, \hat{R}_u)\hat{A}^{1/2}$, where $\hat{R}_u$ is the unstructured correlation matrix estimate (Liang & Zeger, 1986). Specifically,

$$\hat{R}_u = \frac{1}{N}\sum_{i=1}^{N}\hat{A}_i^{-1/2}(y_i - \hat{\pi}_i)(y_i - \hat{\pi}_i)'\hat{A}_i^{-1/2}.$$

Note that $\hat{R}_u$ is obtained without any assumption on the specific structure on the true correlation matrix; in particular, we do not use the estimated working correlation matrix $\hat{R}_W$, which may be specified incorrectly.

According to the asymptotic result in Osius & Rojek (1992), one can argue that G has an approximately normal distribution (as m is bounded and N tends to infinity). The p-value is thus obtained by referring G to a normal distribution with mean Nm and variance $\widehat{\text{var}(G)}$.

Hosmer *et al.* (1997) also reviewed a statistic based on an unweighted sum of residual squares, which surprisingly had a good performance in their simulations (see also Copas, 1989). In our context, we define it to be:

$$U = \sum_{i=1}^{N}\sum_{j=1}^{m}(y_{ij} - \hat{\pi}_{ij})^2 = \hat{\pi}'(1 - \hat{\pi}) + (1 - 2\hat{\pi})'\hat{e} \approx \hat{\pi}'(1 - \hat{\pi}) + (1 - 2\hat{\pi})'(I - H)e.$$

Its mean and variance are approximately

$$\widehat{E(U)} = \hat{\pi}'(1 - \hat{\pi}), \quad \widehat{\text{var}(U)} = (1 - 2\hat{\pi})'(I - \hat{H})\widehat{\text{cov}(Y)}(I - \hat{H}')(1 - 2\hat{\pi}),$$

and $\text{cov}(Y)$ is estimated using either $\widehat{\text{cov}(Y)}_e$ or $\widehat{\text{cov}(Y)}_u$ described above. Again U has an approximately normal distribution.

Remark 1. In the ordinary logistic regression, replacing $\hat{V}$ and $\widehat{\text{cov}(Y)}$ by $\hat{A}$, we obtain the (unconditional) goodness-of-fit tests given in Hosmer *et al.* (1997).

Remark 2. As in le Cessie & Houwelingen (1991) and Hosmer *et al.* (1997), a chi-square distribution can be adopted to approximate the distribution of the non-negative Pearson and the unweighted sum of squares statistics. In our experience, its performance is close to that based on the normal approximation. Hence we will skip its discussion.

Remark 3. The easiest way to use the above two tests is to use the working independence model in GEE; i.e. $R_W = I$ or $V = A$. Our experience from simulation studies showed that using the working independence model yielded results better than those from using the working exchangeable correlation matrix in situations with time-varying covariates and cluster size $m > 2$ (not reported here). The reason is probably that with a more general R_W the more complex form of either the exact result in the appendix or (4) leads to less accurate approximations to the variance of the test statistics when using (and estimating) a more complex working correlation matrix. Hence, from now on, we restrict our discussion of the two tests to that based on the working independence model in GEE.

Remark 4. We emphasize that our tests are proposed for ungrouped binary data, such as when a continuous covariate is present, or, in general, when the number of possible combinations of the covariate values is much larger than the sample size. This corresponds to the so-called increasing-cells asymptotics in Osius & Rojek (1992). If there is a natural partitioning in the covariates, such as when the covariates are all discrete and there are many replications for each combination of the covariate values (i.e. the fixed-cells asymptotics in Osius & Rojek, 1992), Barnhart & Williamson's and Horton *et al*. chi-square tests based on partitioning are more appropriate.

3. Simulations

Simulation studies were conducted to investigate the finite sample performance of the proposed tests. We first use Barnhart & Williamson's Model III to detect an omitted quadratic term. The true model is

$$\text{logit}(\pi_{it}) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2,$$

where $i = 1, \ldots, N$ and $t = 1, 2$; x_i is a continuous covariate generated from a uniform distribution $U(-3, 3)$. The values of βs are determined by $\Pr(y_{it} = 1 | x_i = -1) = 0.2$, $\Pr(y_{it} = 1 | x_i = 3) = 0.95$, and $\Pr(y_{it} = 1 | x_i = -3) = K$ with $K = 0.03$, 0.10, 0.20 and 0.40. In particular, $K = 0.03$ leads to $\beta_2 = 0$, which corresponds to the null hypothesis

$$H_0 : \text{logit}(\pi_{it}) = \beta_0 + \beta_1 x_i.$$

All the simulated data were generated in S-plus, and our computer program was also implemented in S-plus. In particular, we used gee() function to fit a GEE model.

The first scenario we consider is that $\rho = \text{corr}(y_{i1}, y_{i2})$ is constant with all i. Bahadur's (1961) representation was used to generate simulated data. The results with various ρ and N are presented in Table 1. The Pearson chi-square statistic G and the unweighted sum of squares statistic U are based on $\widehat{\text{cov}}(Y)_u$ while the other two, $G2$ and $U2$ are calculated by using $\widehat{\text{cov}}(Y)_e$. We also consider two statistics, G(Indp.) and U(Indp.), obtained under the possibly incorrect assumption that we have independent data.

Under H_0, the test sizes of using G and U are close to the nominal levels, whereas using $G2$ and $U2$ may lead to slightly inflated Type I errors. It may imply that using $\widehat{\text{cov}}(Y)_u$ leads to better performance of the tests. This is not surprising since it is well known that the empirical covariance estimator is inefficient. Between $G2$ and $U2$, it appears that $U2$ is better. Next we compare G and U with the other two statistics, G(Indp.) and U(Indp.). For small ρ, it seems that all four tests have satisfactory size. When $\rho = 0$, using $\hat{R}_u$ does not lose much efficiency as compared to the tests under the correct independence assumption. As ρ increases to 0.5, unsurprisingly, the two tests based on the incorrect independence assumption have inflated

Table 1. *Size (for $K = 0.03$) and power (otherwise) of the goodness-of-fit tests in detecting an omitted quadratic term at the nominal level* 5% (1%) *from* 500 *replications. The within-subject correlation ρ is constant. G or G2 and U or U2 are respectively the Pearson chi-square and the unweighted sum of squares (using the unstructured covariance or the empirical covariance estimator); G*(Indp.) *or U*(Indp.) *is the G or U calculated with $\rho = 0$.*

	$\rho = 0$, $N = 100$				$\rho = 0.2$, $N = 100$				$\rho = 0.5$, $N = 100$				$\rho = 0.5$, $N = 200$			
Statistic	$K = 0.03$	0.10	0.20	0.40	0.03	0.10	0.20	0.40	$K = 0.03$	0.10	0.20	0.40	0.03	0.10	0.20	0.40
G	0.042	0.190	0.498	0.878	0.036	0.174	0.420	0.812	0.036	0.136	0.346	0.712	0.050	0.254	0.632	0.942
	(0.010)	(0.082)	(0.316)	(0.748)	(0.010)	(0.090)	(0.276)	(0.666)	(0.006)	(0.066)	(0.202)	(0.570)	(0.014)	(0.122)	(0.414)	(0.884)
U	0.070	0.172	0.450	0.862	0.050	0.160	0.388	0.808	0.054	0.120	0.326	0.720	0.054	0.216	0.602	0.946
	(0.008)	(0.070)	(0.278)	(0.734)	(0.012)	(0.070)	(0.248)	(0.664)	(0.010)	(0.044)	(0.170)	(0.542)	(0.012)	(0.094)	(0.394)	(0.878)
G2	0.114	0.080	0.274	0.778	0.092	0.078	0.220	0.712	0.082	0.052	0.166	0.584	0.086	0.102	0.458	0.920
	(0.026)	(0.018)	(0.078)	(0.500)	(0.018)	(0.018)	(0.064)	(0.378)	(0.010)	(0.008)	(0.040)	(0.286)	(0.020)	(0.014)	(0.158)	(0.776)
U2	0.086	0.088	0.256	0.802	0.078	0.094	0.248	0.730	0.066	0.078	0.206	0.612	0.064	0.122	0.482	0.918
	(0.020)	(0.022)	(0.104)	(0.524)	(0.014)	(0.026)	(0.090)	(0.420)	(0.010)	(0.016)	(0.054)	(0.324)	(0.016)	(0.024)	(0.202)	(0.788)
G(Indp.)	0.042	0.186	0.504	0.876	0.056	0.202	0.486	0.848	0.092	–	–	–	0.120	–	–	–
	(0.012)	(0.086)	(0.316)	(0.752)	(0.014)	(0.132)	(0.318)	(0.740)	(0.030)	–	–	–	(0.038)	–	–	–
U(Indp.)	0.064	0.168	0.456	0.874	0.076	0.186	0.446	0.842	0.108	–	–	–	0.124	–	–	–
	(0.014)	(0.076)	(0.274)	(0.748)	(0.024)	(0.102)	(0.286)	(0.732)	(0.034)	–	–	–	(0.046)	–	–	–

Type I errors with either $N = 100$ or $N = 200$. In contrast, the two proposed tests (G and U) maintain correct size.

When H_0 does not hold, it appears that the two tests based on G and U are more powerful than those based on $G2$ and $U2$. Although the performances of using the Pearson chi-square statistic G and the unweighted sum of squares statistic U are close, sometimes it seems that using G is slightly more powerful than using U. However, it is possible that the latter is more stable and hence more desirable when some $\hat{\pi}_{it}$ are close to 0 or 1.

An implicit assumption under using $\hat{R}_u$ is that the within-subject correlation structure (parametrized by correlation coefficients) is shared by all subjects; in other words, any $\mathrm{corr}(y_{i1}, y_{i2})$ does not depend on i. To assess the robustness of our tests when this assumption is violated, we generated data with a constant within-subject odds ratio OR as in Barnhart & Williamson (1998) (see also Diggle *et al.*, 1994, p. 150), where

$$\mathrm{OR} = \frac{\Pr(y_{i1} = y_{i2} = 1)\Pr(y_{i1} = y_{i2} = 0)}{\Pr(y_{i1} = 1, y_{i2} = 0)\Pr(y_{i1} = 0, y_{i2} = 1)}$$

does not depend on i. It is easy to verify that in general a constant OR means that the within-subject correlation $\mathrm{corr}(y_{i1}, y_{i2})$ is no longer a constant. From Table 2, it appears that our proposed two tests based on G and U still keep satisfactory size and power. In particular, comparing our results in Table 2 with those in table 3 of Barnhart & Williamson (1998), we feel that the performance of our tests based on G and U is promising. Although the other two test statistics $G2$ and $U2$ do not depend on any assumption on the correlation structure of the response vector, they may still have slightly inflated Type I errors (as in Table 1), which however become closer to the nominal levels as the sample size increases. In particular, the performance of $U2$ appears to be better than that of $G2$.

Now we consider detecting an omitted interaction term. The correct model is

$$\mathrm{logit}(\pi_{it}) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,it} + \beta_3 x_{1,i} x_{2,it},$$

where $i = 1, \ldots, N$ and $t = 1, 2$; $x_{1,i}$ and $x_{2,it}$ are all independently drawn from $U(-2, 2)$; and $\beta_0 = 0$, $\beta_1 = -\beta_2 = 0.5$, and various values of β_3 will be used. The constant within-subject odds ratio is $\mathrm{OR} = 2$. The null model is

$$H_0 : \mathrm{logit}(\pi_{it}) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,it}.$$

The results are presented in Table 3. It can be seen that as β_3 or the sample size N increases, the power of the tests all improve. Again the two tests based on R and U are better than that based on $R2$ or $U2$.

Table 2. *Size* (*for* $K = 0.03$) *and power* (*otherwise*) *of the goodness-of-fit tests in detecting an omitted quadratic term at the nominal level* 5% (1%) *from* 500 *replications. The within-subject odds ratio is constant* (OR = 2)

	$N = 100$				$N = 200$			
Statistic	$K = 0.03$	0.10	0.20	0.40	0.03	0.10	0.20	0.40
G	0.034	0.196	0.472	0.826	0.060	0.314	0.738	0.978
	(0.010)	(0.092)	(0.286)	(0.684)	(0.016)	(0.168)	(0.544)	(0.964)
U	0.060	0.164	0.410	0.822	0.056	0.270	0.698	0.978
	(0.022)	(0.084)	(0.252)	(0.680)	(0.012)	(0.136)	(0.516)	(0.958)
G2	0.090	0.088	0.240	0.718	0.076	0.168	0.586	0.972
	(0.022)	(0.024)	(0.066)	(0.402)	(0.028)	(0.042)	(0.292)	(0.888)
U2	0.074	0.110	0.270	0.732	0.056	0.182	0.592	0.966
	(0.022)	(0.024)	(0.090)	(0.454)	(0.016)	(0.046)	(0.314)	(0.896)

Table 3. *Size (for $\beta_3 = 0$) and power (otherwise) of the goodness-of-fit tests in detecting an omitted two-way interaction term at the nominal level* 5% (1%) *from* 500 *replications. The within-subject odds ratio is constant* (OR = 2)

	$N = 100$				$N = 200$			
Statistic	$\beta_3 = 0$	0.5	1.0	1.5	0	0.5	1.0	1.5
G	0.060	0.340	0.640	0.694	0.042	0.418	0.778	0.812
	(0.016)	(0.184)	(0.530)	(0.604)	(0.002)	(0.294)	(0.694)	(0.744)
U	0.054	0.316	0.636	0.698	0.044	0.402	0.774	0.810
	(0.018)	(0.166)	(0.526)	(0.600)	(0.004)	(0.270)	(0.690)	(0.738)
G2	0.076	0.184	0.584	0.668	0.052	0.346	0.742	0.796
	(0.018)	(0.054)	(0.364)	(0.510)	(0.014)	(0.150)	(0.622)	(0.728)
U2	0.072	0.202	0.582	0.666	0.052	0.338	0.744	0.798
	(0.016)	(0.060)	(0.370)	(0.522)	(0.008)	(0.166)	(0.628)	(0.720)

4. Example

We apply the proposed tests to the Wisconsin epidemiologic study of diabetic retinopathy (Klein *et al.*, 1984) data set, which was also analysed by Barnhart & Williamson (1998). The study goal was to determine the risk factors for diabetic retinopathy. The binary response is the presence of diabetic retinopathy in each of two eyes from each of 720 individuals in the study. As in Barnhart & Williamson (1998), the first model fitted includes four main effects: duration of diabetes, glycosylated hemoglobin level, diastolic blood pressure and body mass index. The Pearson chi-square statistic G is 20505 with mean 1440 and variance 1472992, and the unweighted sum of squares statistic U is 194.6 with mean 213.3 and variance 7.0. Both yield a p-value < 0.0001, indicating the lack-of-fit of the model. Using the statistics $G2$ and $U2$ leads to the p-values 0.1660 and 0.0010 respectively. Based on our observation from the simulation studies that the statistics G and U perform better than $G2$ and $U2$, and that the $U2$ statistic is preferred over $G2$, we conclude that there is an evidence against the current model. Next, we fit a larger model by adding two covariates: the square of duration of diabetes and square of body mass index. The two statistics G and U (with mean and variance) are respectively 1559.8 (1440 and 9387.5) and 183.5 (186.4 and 3.1), leading to the p-values 0.22 and 0.10. Using either of the $G2$ or $U2$ leads to the same conclusion. These results are consistent with those obtained by Barnhart & Williamson (1998).

5. Discussion

In this paper we have proposed two normal-based goodness-of-fit tests for ungrouped correlated binary data. Their performance was investigated through simulation studies and appeared to be satisfactory. The proposed tests are not meant as a replacement of, but complement to, the existing tests, such as Barnhart & Williams's (1998) test. In particular, we emphasize that our proposed tests are intended to be used with ungrouped binary data, where the response is binary and the observed covariates for different subjects (or clusters) are essentially different. When we have grouped binary data, other tests are more appropriate. Although our proposed tests appear to be useful, future studies are warranted to gain more insights on their properties, including their strengths and weaknesses.

To implement the proposed two tests, we need to estimate the covariance matrix of the response vector. Two proposals have been studied. One is the empirical covariance

estimator based on the residuals, as used in the usual sandwich estimator of the estimated regression coefficients in GEE. Another is based on the estimate of the unstructured correlation matrix. The former is less restrictive but also less efficient than the latter. Hence in general we recommend the use of the latter. In addition, although our proposal can accommodate using various working correlation structures in GEE, we find that the working independence model works best. Further studies are needed to investigate how to improve the performance of the tests when other more general working correlation structures are used.

Finally we remark on the limitations of the goodness-of-fit tests. Since all these tests are proposed to detect some general model departures, their power is likely to be limited in practice. Formulating a more specific alternative hypothesis and using a related test can improve the power. For instance, if we suspect that the effect of a covariate is not linear and likely to be quadratic, then directly testing the significance of the adding-in quadratic term will have a higher power than using a general goodness-of-fit test. An attractive point of using a goodness-of-fit test is its convenience. If a goodness-of-fit test rejects the current model, at least it reminds the data analyst that the model is inadequate and some measures have to be taken to fix it. On the other hand, if a goodness-of-fit test does not reject the current model, it does not necessarily mean that the model fits well and the data analyst still needs to use other techniques to confirm the adequacy of the model.

Acknowledgements

The author thanks Dr Huiman Barnhart for providing the WESDR data set. The author is grateful to two referees and the editor for many helpful comments.

References

Bahadur, R. R. (1961). A representation of the joint distribution of responses to *n* dichotomous items. In *Studies in item analysis and prediction*, (ed. H. Solomon), 158–168. Stanford Mathematical Studies in the Social Sciences VI. Stanford University Press, Stanford, CA.

Barnhart, H. X. & Williamson, J. M. (1998). Goodness-of-fit tests for GEE modeling with binary data. *Biometrics* **54**, 720–729.

Copas, J. B. (1989). Unweighted sum of squares test for proportions. *J. Roy. Statist. Soc. Ser. C* **38**, 71–80.

Diggle, P. J., Liang, K.-Y. & Zeger, S. L. (1994). *Analysis of longitudinal data*. Oxford University Press, Oxford.

Horton, N. J., Bebchuk, J. D., Jones, C. L., Lipsitz, S. R., Catalano, P. J., Zahner, G. E. P. & Fitzmaurice, G. M. (1999). Goodness-of-fit for GEE: an example with mental health service utilization. *Statist. Med.* **18**, 213–222.

Hosmer, D. W., Hosmer, T., le Cessie, S. & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statist. Med.* **16**, 965–980.

Hosmer, D. W. & Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Comm. Statist. Theory Methods* **10**, 1043–1069.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. & DeMets, D. L. (1984). The Wisconsin epidemiologic study of diabetic retinopathy: II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years. *Arch. Ophthalmol.* **102**, 520–526.

le Cessie, S. & van Houwelingen, J. C. (1991). A goodness-of-fit test for binary data based on smoothing residuals. *Biometrics* **47**, 1267–1282.

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

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

Osius, G. & Rojek, D. (1992). Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *J. Amer. Statist. Assoc.* **87**, 1145–1152.
Sutradhar, B. C. & Das, K. (1999). On the efficiency of regression estimators in generalized linear models for longitudinal data. *Biometrika* **86**, 459–465.
Tsiatis, A. A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika* **67**, 250–251.
Vonesh, E. F. & Chinchilli, V. M. (1997). *Linear and nonlinear models for the analysis of repeated measurements*. Marcel Dekker, New York.
Zeger, S. L. (1988). The analysis of discrete longitudinal data: commentary. *Statist. Med.* **7**, 161–168.

Received September 2000, in final form May 2001

Wei Pan, Division of Biostatistics, University of Minnesota, A460 Mayo Building, MMC 303, Minneapolis, MN 55455-0378, USA.
E-mail: weip@biostat.umn.edu

Appendix

We derive $\partial S/\partial\beta'$ for any general working correlation matrix R_W being used. We will use some results from matrix differentiation (see, e.g. Vonesh & Chinchilli, 1997, p. 11–16, for a nice introduction). We need first define some matrix operators. For any matrix B, vec(B) creates another column vector by simply stacking the columns of B one by one. For any $r \times s$ matrix $A = (a_{ij})$ and any matrix B, define

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

And we use I_r to denote a $r \times r$ identity matrix.

Using (1) we have

$$\begin{aligned} \frac{\partial S}{\partial\beta'} &= \sum_{i=1}^{N} \frac{\partial[x_i' A_i^{1/2} R_W^{-1} A_i^{-1/2}(y_i - \pi_i)]}{\partial\beta'} \\ &= \sum_{i=1}^{N} \left(A_i^{-1/2}(y_i - \pi_i) \otimes I_k\right)' \frac{\partial \operatorname{vec}(x_i' A_i^{1/2} R_W^{-1})}{\partial\beta'} \\ &\quad + \left(I_1 \otimes x_i' A_i^{1/2} R_W^{-1}\right) \frac{\partial[A_i^{-1/2}(y_i - \pi_i)]}{\partial\beta'}. \end{aligned}$$

Since

$$\begin{aligned} \frac{\partial \operatorname{vec}(x_i' A_i^{1/2} R_W^{-1})}{\partial\beta'} &= \left(I_{N_i} \otimes x_i'\right) \frac{\partial \operatorname{vec}(A_i^{1/2} R_W^{-1})}{\partial\beta'} = \left(I_{N_i} \otimes x_i'\right) \left[(R_W^{-1} \otimes I_{N_i})' \frac{\partial \operatorname{vec}(A_i^{1/2})}{\partial\beta'}\right] \\ &= \left(R_W^{-1} \otimes x_i'\right) \frac{\partial \operatorname{vec}(A_i^{1/2})}{\partial\beta'}, \end{aligned}$$

and

$$\frac{\partial[A_i^{-1/2}(y_i - \pi_i)]}{\partial\beta'} = [(y_i - \pi_i) \otimes I_{N_i}]' \frac{\partial \operatorname{vec}(A_i^{-1/2})}{\partial\beta'} + (I_1 \otimes A_i^{-1/2}) \frac{\partial(y_i - \pi_i)}{\partial\beta'},$$

using $\partial\pi_i/\partial\beta' = A_i x_i$, we obtain

$$\begin{aligned}\partial S/\partial\beta' &= \sum_{i=1}^{N}\left\{[(y_i-\pi_i)'A_i^{-1/2}R_W^{-1}\otimes x_i']\frac{\partial\ \mathrm{vec}(A_i^{1/2})}{\partial\beta'}\right.\\ &\qquad\left.+[(y_i-\pi_i)'\otimes x_i'A_i^{1/2}R_W^{-1}]\frac{\partial\ \mathrm{vec}(A_i^{-1/2})}{\partial\beta'}-x_i'A_i^{1/2}R_W^{-1}A_i^{1/2}x_i\right\}\\ &= \sum_{i=1}^{N}\left\{\sum_{j=1}^{N_i}\frac{1}{2}b_{ij}v_{ij}^{1/2}(1-2\pi_{ij})x_{ij}x_{ij}'\right.\\ &\qquad\left.-\sum_{j=1}^{N_i}\frac{1}{2}e_{ij}v_{ij}^{-1/2}(1-2\pi_{ij})(x_i'A_i^{1/2}R_W^{-1})_jx_{ij}'-x_i'A_iV_i^{-1}A_ix_i\right\},\end{aligned}$$

where b_{ij} is the jth element of $(y_i-\pi_i)'A_i^{-1/2}R_W^{-1}$, $v_{ij}=\pi_{ij}(1-\pi_{ij})$, and $(x_i'A_i^{1/2}R_W^{-1})_j$ is the jth column of $x_i'A_i^{1/2}R_W^{-1}$.