BIOMETRICS 54, 720–729
June 1998

Goodness-of-Fit Tests for GEE Modeling with Binary Responses

Huiman X. Barnhart*

Department of Biostatistics, The Rollins School of Public Health of Emory University,
1518 Clifton Road, NE, Atlanta, Georgia 30322, U.S.A.

and

John M. Williamson

Division of HIV/AIDS Prevention, Surveillance and Epidemiology (E-48),
Centers for Disease Control and Prevention,
1600 Clifton Road, NE, Atlanta, Georgia 30333, U.S.A.

SUMMARY

Analysis of data with repeated measures is often accomplished through the use of generalized estimating equations (GEE) methodology. Although methods exist for assessing the adequacy of the fitted models for uncorrelated data with likelihood methods, it is not appropriate to use these methods for models fitted with GEE methodology. We propose model-based and robust (empirically corrected) goodness-of-fit tests for GEE modeling with binary responses based on partitioning the space of covariates into distinct regions and forming score statistics that are asymptotically distributed as chi-square random variables with the appropriate degrees of freedom. The null distribution and the statistical power of the proposed goodness-of-fit tests were assessed using simulated data. The proposed goodness-of-fit tests are illustrated by two examples using data from clinical studies.

1. Introduction

Studies involving binary outcomes are quite common in the health sciences. The logistic regression model is a widely used method for analyzing such data, and methods exist for assessing the adequacy of the fitted model (Tsiatis, 1980; Hosmer and Lemeshow, 1989, Chapter 5; le Cessie and van Houwelingen, 1991; for comparisons of seven goodness-of-fit tests for ordinary logistic regression models, see Hosmer et al. (1997)). However, the analysis of correlated binary responses is often accomplished through the use of generalized estimating equations (GEE) methodology (Liang and Zeger, 1986; Zeger and Liang, 1986) for parameter estimation. Assessment of the adequacy of the fitted GEE model is problematic since no likelihood exists and the residuals (observed minus expected terms) are correlated within a cluster. We propose model-based and robust (empirically corrected) goodness-of-fit tests based on partitioning the space of covariates into distinct regions (similar to Tsiatis (1980) and Lipsitz and Buoncristiani (1994)) and forming score statistics that are asymptotically distributed as chi-square random variables with the appropriate degrees of freedom.

In Section 2, we present the proposed goodness-of-fit test statistics and derive their asymptotic distributions. Simulation results presented in Section 3 evaluate the asymptotic distributions of the goodness-of-fit test statistics and the power of the goodness-of-fit tests for different alternatives to the null hypothesis. Two examples using data from clinical studies illustrate the use of the proposed goodness-of-fit tests in Section 4. We conclude with a discussion in Section 5.

* *Corresponding author's email address:* hbarnha@sph.emory.edu

Key words: Correlated data; GEE modeling; Goodness-of-fit test; Logistic regression; Score test.

2. Goodness-of-Fit Test

Suppose there is a random sample of N individuals, where the ith individual has T_i binary measurements (0 or 1), $i = 1, \ldots, N$. For simplicity, we assume $T_i \equiv T$ for all i. Let $\mathbf{Y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$ be a $T \times 1$ vector consisting of binary responses and $\mathbf{X}_i = [\mathbf{1}, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iP}]$ the $T \times (P+1)$ design matrix for the ith subject ($i = 1, \ldots, N$), where $\mathbf{1}$ is a $T \times 1$ vector comprised of ones. The covariates may be either time independent or time dependent. The usual GEE modeling for binary outcomes has the following setting:

$$\text{logit}(\boldsymbol{\mu}_i) = \mathbf{X}_i \boldsymbol{\beta}, \tag{1}$$

where $\boldsymbol{\mu}_i = \text{E}(\mathbf{Y}_i)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_P)'$, and $\text{logit}(a) = log(a/(1-a))$. Estimation of $\boldsymbol{\beta}$ is obtained by solving the generalized estimating equations (Liang and Zeger, 1986; Zeger and Liang, 1986)

$$\sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \beta_p} \right)' \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0, \qquad p = 1, \ldots, P+1, \tag{2}$$

with $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i \mathbf{A}_i^{1/2}, \mathbf{A}_i = \text{diag}(\text{var}(y_{i1}), \ldots, \text{var}(y_{iT}))$, where $\text{var}(y_{it}) = \mu_{it}(1 - \mu_{it})$ is the variance of y_{it} and $\mathbf{R}_i$ is the working correlation matrix for $\mathbf{Y}_i$. Note that $\text{diag}(\mathbf{B}_1, \ldots, \mathbf{B}_k)$ denotes the block diagonal matrix with $\mathbf{B}_1, \ldots, \mathbf{B}_k$ as the diagonal entries, where $\mathbf{B}_i$ can be either a matrix or a scalar.

We describe the proposed goodness-of-fit statistics by first partitioning the covariate space $\mathbf{X} = (x_1, \ldots, x_P)'$ into M distinct regions in P-dimensional space. Let $\mathbf{I}_{it} = (I_{it1}, \ldots, I_{itM})'$ be an $M \times 1$ vector, where I_{itm} is the indicator variable that equals one if the ith subject is in the mth region at the tth occasion and zero otherwise. We define the $T \times M$ matrix $\mathbf{I}_i$ as

$$\mathbf{I_i} = [\mathbf{I}_{i1}, \ldots, \mathbf{I}_{iT}]'. \tag{3}$$

Let $\mathbf{Z}_T$ be the $T \times (T-1)$ matrix where the first row has entries zero and the remaining $(T-1)$ rows form a $(T-1) \times (T-1)$ identity matrix. Consider the model

$$\text{logit}(\boldsymbol{\mu}_i) = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_T \boldsymbol{\tau} + \mathbf{I}_i \boldsymbol{\gamma} + \mathbf{S}_i \boldsymbol{\rho}, \tag{4}$$

where $\mathbf{S}_i = [\mathbf{0}, \text{diag}(\mathbf{I}_{i2}, \ldots, \mathbf{I}_{iT})]'$ is a $T \times (T-1)M$ matrix and $\mathbf{0}$ is a $(T-1)M \times 1$ vector of zeros. Note that $\boldsymbol{\tau}$ is the $(T-1) \times 1$ vector of time effects (the first occasion is the reference time point), $\boldsymbol{\gamma}$ is the $M \times 1$ vector of region effects, and $\boldsymbol{\rho}$ is the $(T-1)M \times 1$ vector of time and region interaction effects because each column of $\mathbf{S}_i$ results from componentwise multiplication of two column vectors, one column vector from $\mathbf{Z}_T$ and the other from $\mathbf{I}_i$. A goodness-of-fit statistic consists of testing H_0: $\boldsymbol{\theta} = 0$, where $\boldsymbol{\theta} = [\boldsymbol{\tau}', \boldsymbol{\gamma}', \boldsymbol{\rho}']'$ is a $J \times 1$ vector with $J = (T-1) + M + (T-1)M$.

Let $L = P + 1 + J$ be the number of parameters in the model presented in (4). Denote $\mathbf{U}$ be the $L \times 1$ vector with lth component

$$U_l = \sum_{i=1}^{N} \hat{\mathbf{D}}_{il}' \hat{\mathbf{V}}_i^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i) \tag{5}$$

for $l = 1, \ldots, L$, where $\hat{\mathbf{D}}_{il} = \partial \hat{\boldsymbol{\mu}}_i / \partial \beta_l$ for $l \leq P+1$, $\hat{\mathbf{D}}_{il} = \partial \hat{\boldsymbol{\mu}}_i / \partial \theta_{l-P-1}$ for $l > P+1$, $\hat{\boldsymbol{\mu}}_i = \text{logit}^{-1}(\mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{Z}_T \boldsymbol{\tau} + \mathbf{I}_i \boldsymbol{\gamma} + \mathbf{S}_i \boldsymbol{\rho})$, and $\hat{\boldsymbol{\beta}}$ is obtained as the solution to equation (2). Then, under H_0: $\boldsymbol{\theta} = 0$, the asymptotic distribution of $\mathbf{U}$ is multivariate normal with mean zero and covariance matrix (Liang and Zeger, 1986)

$$\mathbf{W}_R = \sum_{i=1}^{N} \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \text{cov}(\mathbf{Y}_i) \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i, \tag{6}$$

where $\hat{\mathbf{D}}_i = [\hat{\mathbf{D}}_{i1}, \ldots, \hat{\mathbf{D}}_{iL}]$ is a $T \times L$ matrix. Note that $\text{cov}(\mathbf{Y}_i)$ can be consistently estimated by $(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)'$ (Liang and Zeger, 1986). If the correlation matrix $\mathbf{R}_i$ is correctly specified, then the asymptotic covariance matrix of $\mathbf{U}$ reduces to $\mathbf{W} = \Sigma_{i=1}^{N} \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i$.

Let

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix} \qquad \mathbf{W}_R = \begin{pmatrix} \mathbf{A}_R & \mathbf{B}_R' \\ \mathbf{B}_R & \mathbf{C}_R \end{pmatrix} \qquad \mathbf{W} = \begin{pmatrix} \mathbf{A} & \mathbf{B}' \\ \mathbf{B} & \mathbf{C} \end{pmatrix}$$

be the partitioning for $\mathbf{U}, \mathbf{W}_R$, and $\mathbf{W}$, where $\mathbf{U}_2$ is the $J \times 1$ vector and $\mathbf{C}_R$ and $\mathbf{C}$ are $J \times J$ matrices. Under H_0: $\boldsymbol{\theta} = 0$, both the proposed robust (empirically corrected) goodness-of-fit test statistic

$$Q_R = \mathbf{U}_2'(\mathbf{C}_R - \mathbf{B}_R\mathbf{A}_R^{-1}\mathbf{B}_R')^{-}\mathbf{U}_2$$

and the proposed model-based goodness-of-fit test statistic

$$Q = \mathbf{U}_2'(\mathbf{C} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-}\mathbf{U}_2$$

are asymptotically distributed as chi-square random variables with

$$\text{d.f.} = \text{rank}((\mathbf{C}_R - \mathbf{B}_R\mathbf{A}_R^{-1}\mathbf{B}_R')^{-}) = \text{rank}((\mathbf{C} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-}),$$

where $\mathbf{G}^{-}$ is any generalized inverse of the matrix $\mathbf{G}$. The degrees-of-freedom for the chi-square random variables do not equal the number of parameters in $\boldsymbol{\theta}$ because of linear dependencies between the covariates in the model and the covariates from the region partitioning, i.e., $(\mathbf{C}_R - \mathbf{B}_R\mathbf{A}_R^{-1}\mathbf{B}_R')$ and $(\mathbf{C} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}')$ are singular matrices. Let $\mathbf{H}_1$ and $\mathbf{H}_2$ be the design matrices in models (1) and (4), respectively. Then intuitively, the degrees-of-freedom of the above chi-square random variables is equal to rank$(\mathbf{H}_2)$ − rank$(\mathbf{H}_1)$. Let $\mathbf{H}_{2i} = \{h_{itj}\} = [\mathbf{X}_i, \mathbf{Z}_T, \mathbf{I}_i, \mathbf{S}_i]$ be the $T \times (P + 1 + J)$ design matrix for the ith subject in model (4). It can be easily shown that the tjth element of $\hat{\mathbf{D}}_i$ is equal to $\hat{\mu}_{it}(1 - \hat{\mu}_{it})h_{itj}$. Therefore, the goodness-of-fit test statistics Q and Q_R can be readily calculated once $\hat{\boldsymbol{\beta}}$ is obtained from the estimating equations (2).

3. Simulations

In order to assess the performance of the proposed goodness-of-fit tests, we used data simulated with known distributions from models in the alternative hypotheses to test the goodness-of-fit of three proposed models. The null and alternative hypotheses of the three models are

$$\begin{aligned}
\text{Model I: } &\text{H}_0\text{: logit}(\mu_{it}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{it2}\\
&\text{H}_1\text{: logit}(\mu_{it}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{it2} + \beta_3 x_{i1} x_{it2},\\
\text{Model II: } &\text{H}_0\text{: logit}(\mu_{it}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{t2} + \beta_3 x_{i3}\\
&\text{H}_1\text{: logit}(\mu_{it}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{t2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i3},\\
\text{Model III: } &\text{H}_0\text{: logit}(\mu_{it}) = \beta_0 + \beta_1 x_i\\
&\text{H}_1\text{: logit}(\mu_{it}) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2,
\end{aligned}$$

for $i = 1, \ldots, N = 400$ denoting subject and $t = 1, \ldots, T = 2$ denoting occasion. For Model I, x_1 is a time-independent binary covariate and x_2 is a time-dependent continuous covariate. For Model II, x_1 is again a time-independent binary covariate, x_2 is an indicator variable for the second occasion, and x_3 is a time-independent continuous covariate. For Model III, x is a time-independent continuous covariate. Note that both Models I and II test for interaction departure and Model III tests for quadratic departure.

For each model, a total of 500 data sets each containing 400 individuals at two time points ($t = 1, 2$) were generated under the alternative hypotheses. The correlation between an individual's responses was induced by assuming that the true model has an odds ratio of 2.0 between the responses at the two occasions.

For Model I, 200 observations were generated for each of the two covariate levels of x_1, 0 and 1. At each occasion, the second covariate, x_2, was generated as a uniform$(-1, 1)$ random variable. The values of $\boldsymbol{\beta}$ used to generate the data were $\beta_0 = 0.0$, $\beta_1 = -0.5$, and $\beta_2 = 0.5$. Eleven sets of simulations were conducted as β_3 ranged from 0.0 to 1.5 by increments of 0.15. For Model II, 200 observations were again generated for each of the two covariate levels of x_1, 0 and 1. The third covariate, x_3, was generated as a uniform$(-1, 1)$ random variable with the same value at each occasion. The values of $\boldsymbol{\beta}$ used to generate the data were $\beta_0 = 0.0$, $\beta_1 = -0.5$, $\beta_2 = 0.5$, and $\beta_3 = 0.5$. Again, eleven sets of simulations were conducted as β_4 ranged from 0.0 to 1.5 by increments of 0.15.

For Model III, the covariate x was generated as a uniform$(-3, 3)$ random variable with the same value at each occasion. The values of $\boldsymbol{\beta}$ used to generate the data were determined by $\Pr(Y_{it} = 1 \mid x = -1) = \mu_{it}(-1) = 0.2, \Pr(Y_{it} = 1 \mid x = 3) = \mu_{it}(3) = 0.95$, and $\Pr(Y_{it} = 1 \mid x = -3) = \mu_{it}(-3) = K$, with $K = 0.05, 0.10, 0.15, 0.2, 0.3$, and 0.4. This scheme generated models where the quadratic departure in the logit function becomes progressively more pronounced (see Figure 1). The corresponding $\boldsymbol{\beta}$ values are displayed in Table 1.

To conduct the proposed goodness-of-fit tests, the following covariate partitioning was utilized. For Model I, the regions were partitioned as $(x_1 = 0, x_2 \leq 0)$, $(x_1 = 0, x_2 > 0)$, $(x_1 = 1, x_2 \leq 0)$,

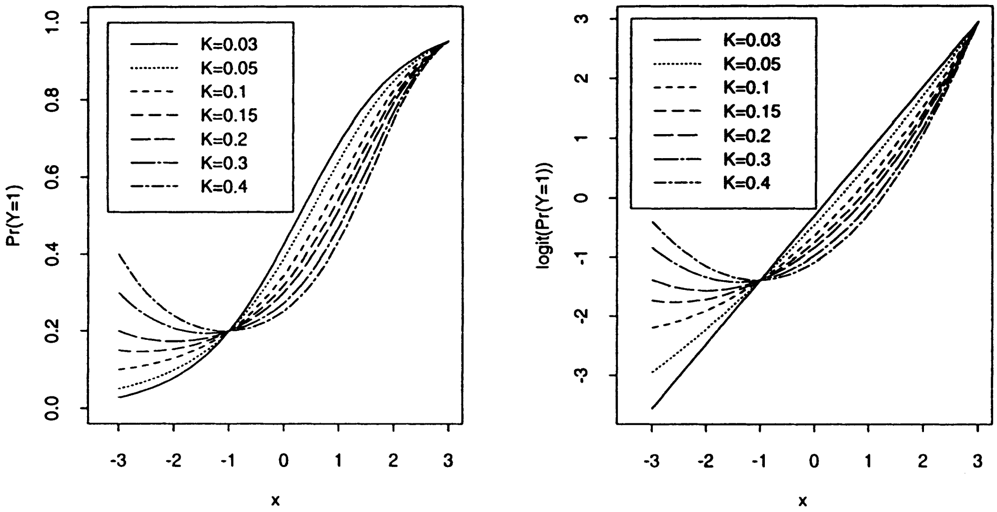

Figure 1. $\Pr(Y = 1)$ and $\operatorname{logit}(\Pr(Y = 1))$ for Model III.

$(x_1 = 1, x_2 > 0)$ and $\boldsymbol{\theta}$ is a 9×1 vector consisting of a scalar τ and the 4×1 vectors γ and ρ. The region partitions for Model II were the same as for Model I except that x_2 is replaced by x_3. For this situation, $\boldsymbol{\theta}$ is also a 9×1 vector. For Model III, the regions were partitioned as $-3 \le x \le -1, -1 < x \le 1, 1 < x \le 3$ and $\boldsymbol{\theta}$ is a 7×1 vector. The goodness-of-fit test statistics have five degrees of freedom. Note that, for Model I, the same subject can belong to different covariate regions at the two occasions, but each subject belongs to the same covariate region at each occasion for Models II and III.

For each of the three models, 500 data sets, and values of the parameter in the alternative hypothesis, the proposed goodness-of-fit test statistics were calculated. The data sets were first analyzed with the working correlation matrix misspecified as an identity matrix, and then the model-based goodness-of-fit test statistic (Q_m) and the robust goodness-of-fit test statistic (Q_{Rm}) were calculated. The data sets were then reanalyzed with the correctly specified working correlation matrix, i.e., one with the correlation calculated so that there was an odds ratio of 2.0 between the two responses of a subject. Again, both the model-based goodness-of-fit test statistic (Q_c) and the robust goodness-of-fit test statistic (Q_{Rc}) were calculated. The power of each test statistic was estimated as the number of data sets out of 500 in which the goodness-of-fit test was significant at the $\alpha = 0.05$ significance level. This estimate approximates the probability of rejecting the null hypothesis given that the alternative is true. Table 2 presents the power calculations for the four goodness-of-fit tests for Models I and II, and Table 3 presents the power estimates for Model III.

Table 2 reveals that the power for detecting interaction departure is quite low. For example, when the interaction parameter is 0.45, which is about the same magnitude as the main effect parameter (i.e., 0.5), the statistical power for both models ranges from 13 to 18%. However, the power for detecting quadratic departure is high (see Table 3). For example, when the quadratic

Table 1
Parameters used in simulating data for Model III

	Parameters		
$K = \Pr(Y = 1 \mid x = -3)$	β_0	β_1	β_2
0.03	−0.30361100	1.08268300	0.0
0.05	−0.45541654	0.98147966	0.05060184
0.10	−0.64222010	0.85694390	0.11286970
0.15	−0.75787600	0.77984000	0.15142170
0.20	−0.84495270	0.72178890	0.18044720
0.30	−0.97970180	0.63195610	0.22536360
0.40	−1.09016000	0.55831730	0.26218300

Table 2

Power of the proposed goodness-of-fit tests for detecting interaction departure[a,b]

Interaction parameter	Model I				Model II			
	Q_m	Q_{Rm}	Q_c	Q_{Rc}	Q_m	Q_{Rm}	Q_c	Q_{Rc}
0.00	3.6	4.4	4.4	4.4	4.2	4.8	4.6	4.6
0.15	3.6	5.0	4.8	4.6	6.2	6.0	5.8	6.0
0.30	7.4	8.8	8.6	8.8	11.0	11.0	10.2	11.0
0.45	15.4	18.0	16.4	17.2	15.4	13.4	14.6	13.4
0.60	23.6	25.2	25.2	26.8	24.6	23.4	23.0	23.8
0.75	39.6	41.6	41.6	42.8	40.0	35.2	34.0	35.2
0.90	54.0	55.8	56.6	57.6	54.4	50.6	50.2	51.0
1.05	70.6	74.4	73.4	75.0	68.8	62.8	63.4	63.4
1.20	81.2	82.4	82.6	83.2	83.0	79.4	79.0	79.6
1.35	90.0	90.4	89.6	89.6	91.0	88.2	88.4	89.0
1.50	95.2	95.2	96.4	96.4	94.8	94.2	94.0	94.4

[a] The power estimates are based on 500 data sets with a sample size of 400.

[b] Q_m, Q_{Rm}, Q_c, and Q_{Rc} refer to misspecified model-based, misspecified robust, correctly specified model-based, and correctly specified robust goodness-of-fit test statistics, respectively.

parameter is about 0.26 (corresponding to $K = 0.4$), which is relatively small compared to the main effect parameter β_1 (= 0.56) (see Table 1), the power of the goodness-of-fit tests is about 99% for a sample size of 400. Noticing the high power for detecting quadratic departure with a sample size of 400, we further computed the power for detecting quadratic departure via simulation using smaller sample sizes (200 and 100; see Table 3). For $K = 0.4$, the statistical power is 82% (52%) for data sets with a sample size of 200 (100). Similar low power for detecting interaction departure and high power for detecting quadratic departure were also observed in the goodness-of-fit tests for ordinary logistic regression models with independent binary responses (Hosmer et al., 1997).

When the parameter associated with the interaction covariate (for Models I and II) or with the quadratic term (for Model III) is zero, the proposed goodness-of-fit tests have approximately the specified 5% Type I error rate for a sample size of 400 (and 200 in Model III), i.e., the nominal significance level of $\alpha = 0.05$ (first row of Tables 2 and 3). However, the Type I error rate is inflated for the robust goodness-of-fit test when the sample size is equal to 100 (see columns 11 and 13 of the first row in Table 3). Nevertheless, the Type I error rate is close to 0.05 for the model-based goodness-of-fit test (with misspecified or correctly specified working correlation matrix) in Table 3. The inflated Type I error rate of the robust test is probably due to small frequencies of $Y = 1$ in some partitioned regions. Emrich and Piedmonte (1992) reported that, for small samples, the test of regression parameters equal to zero has correct Type I error rate when the model-based

Table 3

Power of the proposed goodness-of-fit tests for detecting quadratic departure (Model III)[a,b]

	Sample size											
	400				200				100			
K^c	Q_m	Q_{Rm}	Q_c	Q_{Rc}	Q_m	Q_{Rm}	Q_c	Q_{Rc}	Q_m	Q_{Rm}	Q_c	Q_{Rc}
0.03	4.6	7.0	5.6	7.0	3.2	7.6	3.0	7.6	4.8	18.0	5.0	17.0
0.05	9.0	10.0	8.6	9.8	7.4	9.4	8.0	10.0	5.4	15.0	5.4	15.8
0.10	34.8	33.6	32.2	34.2	12.8	12.8	13.2	13.2	11.8	14.6	10.6	14.6
0.15	59.6	55.6	56.2	57.0	35.0	31.0	31.4	31.8	16.0	18.2	15.2	18.2
0.20	82.0	78.2	80.0	79.4	49.4	48.0	47.8	49.2	20.8	24.2	19.0	25.0
0.30	94.6	93.4	93.6	93.8	69.8	66.8	66.4	67.8	40.2	40.4	36.2	40.2
0.40	99.6	99.4	99.4	99.4	84.6	81.6	81.2	82.6	53.2	53.8	48.4	54.6

[a] Power estimates are based on 500 data sets.

[b] Q_m, Q_{Rm}, Q_c, and Q_{Rc} refer to misspecified model-based, misspecified robust, correctly specified model-based, and correctly specified robust goodness-of-fit test statistics, respectively.

[c] $K = \Pr(Y = 1 \mid x = -3)$. See β parameters in Table 1.

covariance matrix and true correlation matrix were used. However, the test has inflated Type I error rate when the robust covariance matrix was employed (using either correctly specified or misspecified correlation matrix). In our situation where $\beta_3 = 0$ in Model III, the $\Pr(Y = 1)$ is small (ranging from 0.03 to 0.2) when $-3 \leq x \leq -1$. This implies that, when the sample size is 100, one might expect $33 \times 0.03 \approx 1$ to $33 \times 0.2 \approx 7$ subjects with an x value in $[-3, -1]$ and the response $Y = 1$. The low frequency of $Y = 1$ might cause difficulty in estimation of the robust variance corresponding to the parameter for region $-3 \leq x \leq -1$. We observed that Q_R was relatively large compared to Q in about 65 of the 500 simulations. In almost in all cases, the frequency of $Y = 1$ in region $-3 \leq x \leq -1$ was either zero or one for the simulated data sets associated with large values of Q_R. This finding may explain why the Type I error rate is inflated.

For sample sizes of 200 or 400, the power of both robust goodness-of-fit tests (with misspecified and correctly specified working correlation matrices) is very similar to the power of the model-based goodness-of-fit test with correctly specified working correlation matrix. However, the power of the model-based goodness-of-fit test with misspecified working correlation matrix does not agree closely with the power of the other tests. The power of the robust goodness-of-fit test may be inflated for small sample sizes because some partitioned regions contain low frequencies of $Y = 1$ or $Y = 0$.

4. Examples

Two examples are presented to illustrate use of the proposed goodness-of-fit tests. Since we do not know the true correlation matrix, an identity correlation matrix was used in computing goodness-of-fit statistics in all examples. The first illustrative example uses data presented by Agresti (1990). A longitudinal study was conducted to compare a new drug and a standard drug for treatment of subjects suffering mental depression. Subject's degree of depression was diagnosed as mild or severe at baseline and a total of 340 subjects were randomized to one of the two treatments. The subjects were followed at 1, 2, and 4 weeks after the treatment allocation. Each subject was classified as normal or abnormal in terms of suffering from mental depression at each follow-up. The data are displayed in Table 4.

We fit two models to the data. The first model only includes the main effects of time, diagnosis and treatment, and the second model includes the same main effects and the treatment and time interaction. Because all the covariates are discrete, the covariate categories were used to form four regions with frequencies displayed in the last column of Table 4. The robust parameter estimates and goodness-of-fit tests are displayed in Table 5.

Both the goodness-of-fit tests suggest that the model with only main effects did not fit the data well ($Q = 33.83, Q_R = 38.93; p < 0.001$). There is a significant time and treatment interaction effect ($\beta_4 = -1.02; p < 0.001$), indicating that patients with new drug treatment improved significantly faster than the patients with the standard drug. The model with this interaction term included has a good fit to the data ($Q = 4.45, Q_R = 4.57; p = 0.73, p = 0.71$). The parameter estimates and the goodness-of-fit tests obtained here are very similar to the results obtained using a weighted least squares approach (goodness-of-fit statistic = 4.15, d.f. = 7; $p = 0.76$; Agresti, 1990). Thus, the proposed goodness-of-fit tests successfully detected the interaction departure.

The above example deals with data with few discrete covariates. In practice, statisticians often need to analyze data with a number of discrete and continuous covariates. We present a second example to illustrate the use of the proposed goodness-of-fit tests in this situation. The data set analyzed is from the Wisconsin Epidemiological Study of Diabetic Retinopathy (Klein et al., 1984). A total of 996 insulin-taking younger-onset diabetics in southern Wisconsin were examined for severity of diabetic retinopathy. The goal of the study is to determine the risk factors for diabetic

Table 4
Cross-classification of responses at three times (N = normal, A = abnormal) by diagnosis and treatment

		Response at 1, 2, and 4 weeks								
Diagnosis	Treatment	NNN	NNA	NAN	NAA	ANN	ANA	AAN	AAA	Total
Mild	Standard	16	13	9	3	14	4	15	6	80
Mild	New drug	31	0	6	0	22	2	9	0	70
Severe	Standard	2	2	8	9	9	15	27	28	100
Severe	New drug	7	2	5	2	31	5	32	6	90

Table 5
Parameter estimates and goodness-of-fit tests for the data from the depression study

	Model 1			Model 2		
Covariates	Estimate	S.E.	p-value	Estimate	S.E.	p-value
Intercept	−0.88	0.16	<0.001	−1.40	0.18	<0.001
Time (0, 1, 2)[a]	0.90	0.09	<0.001	1.50	0.14	<0.001
Diagnosis ($1 = \text{severe}, 0 = \text{mild}$)	1.29	0.14	<0.001	1.31	0.15	<0.001
Treatment ($1 = \text{newdrug}, 0 = \text{standard}$)	−0.88	0.14	<0.001	0.06	0.23	0.79
Treatment × time interaction	—	—	—	−1.02	0.19	<0.001

Goodness-of-Fit

	Statistic	d.f.	p-value	Statistic	d.f.	p-value
Q	33.83	8	<0.001	4.45	7	0.73
Q_R	38.93	8	<0.001	4.57	7	0.71

[a] The values for time, 0, 1, 2, correspond to the logarithm to the base 2 of weeks 1, 2, and 4.

retinopathy. The severity of diabetic retinopathy was originally graded on a 10-point scale. For the purpose of illustration, the 10-point scale was combined to form two categories: absence and presence of diabetic retinopathy. The cross-classification of diabetic retinopathy in left and right eyes for the 720 individuals with complete information is presented in Table 6. In this example, the cluster is the individual with correlated binary responses at different eyes.

Variable selection was performed in the following manner. First, a univariate analysis was performed where a logit model with only 1 of the 13 covariates was fit to the data. The proposed goodness-of-fit tests were also performed along with each logit model. To better detect quadratic and interaction departure, the regions were formed by using categories for the discrete covariate and the quartiles for the continuous covariate. Results of the univariate analysis are presented in Table 7.

If the proposed goodness-of-fit tests were statistically significant at the $\alpha = 0.2$ level, we then considered fitting the logit model with an additional quadratic term or an interaction term of the covariate with a variable indicating eye ($0 =$ left eye, $1 =$ right eye). We found that the quadratic terms of the two covariates duration of diabetes and body mass index are statistically significant and the inclusion of these terms improved the fit of the model. Next, we fit a logit model with the significant marginal and quadratic covariates found in the univariate analysis. In this process, we found that the four covariates intraocular pressure, systolic blood pressure, proteinuria, and pulse rate are no longer statistically significant. We derived the final model after removing these four covariates (Table 8).

Note that there are four continuous covariates in the final model. If we use the quantiles to partition the covariate space, there will be 256 regions. There will likely be empty cells with such a large number of regions. From the simulation results, we know that Q_R is not stable when there is a small cell count in a region. To deal with this problem, we used the medians instead of the quantiles of the four continuous covariates to form 16 regions. Together with the variable indicating eye ($0 =$ left eye, $1 =$ right eye) and the interaction of region and eye effects, we have added a total of 33 parameters in the alternative model for calculating the goodness-of-fit statistics. We found that the final model fits the data well ($p = 0.37$ or 0.17; Table 8). We also fit the model without the quadratic terms and found this model has a poor fit ($p < 0.001$). Therefore, the goodness-of-

Table 6
Frequency of diabetic retinopathy

	Left eye		
Right eye	Absence	Presence	Total
Absence	237	38	275
Presence	31	414	445
Total	268	452	720

Table 7

Univariate analysis of the diabetic retinopathy study

Covariate[a]	Parameter estimate	Robust p-value	Q_R	d.f.	p-value
Eye-specific covariate					
Refractive error	−0.020	0.51	23.3	7	0.001
Intraocular pressure	0.062	0.002	7.6	7	0.37
Person-specific covariate					
Age at diagnosis of diabetes (years)	0.011	0.25	5.9	7	0.55
Duration of diabetes (years)	0.243	<0.001	45.7	7	<0.001
Glycosylated hemoglobin level	0.073	0.013	18.0	7	0.012
Systolic blood pressure	0.035	<0.001	2.6	7	0.92
Diastolic blood pressure	0.047	<0.001	8.0	7	0.33
Body mass index	0.715	<0.001	21.2	7	0.003
Pulse rate (beats/30 seconds)	0.032	0.006	4.0	7	0.78
Gender (male $= 0$, female $= 1$)	0.145	0.32	1.3	2	0.52
Proteinuria (absent $= 0$, present $= 1$)	1.344	<0.001	0.9	2	0.64
Doses of insulin per day (0, 1, 2)	−0.235	0.11	7.0	3	0.07
Residence (urban $= 0$, rural $= 1$)	0.043	0.76	0.8	2	0.67

[a] One covariate, macular edema (absent = 0, present = 1), was not used in the model fitting because individuals with macular edema all had diabetic retinopathy.

fit tests are able to detect the observed quadratic departure. The final model implies that higher glycosylated hemoglobin level and higher diastolic blood pressure are associated with the presence of diabetic retinopathy. The duration of diabetes and body mass index are significantly associated with the presence of diabetic retinopathy in a quadratic pattern, where the probability of diabetic retinopathy behaves like a bell shape as duration of diabetes or body mass index increases.

5. Discussion

We have proposed model-based and robust goodness-of-fit tests for GEE modeling with binary responses. The proposed tests have high power for detecting quadratic departure but have low power for detecting interaction departure. The robust goodness-of-fit test performs better than the model-based goodness-of-fit test when the working correlation matrix is misspecified and the sample size is large. When the sample size is small, the model-based goodness-of-fit performs better in terms of Type I error rate. Two examples using data from clinical studies illustrated the use of the proposed goodness-of-fit tests.

Note that, if the number of time points and/or covariates is large, the number of additional parameters in equation (4) can explode. Furthermore, if the number of covariates is large, the

Table 8

Final GEE logistic regression model for diabetic retinopathy study

	Model without quadratic terms			Final model		
Covariates	Estimate	S.E.	p-value	Estimate	S.E.	p-value
Intercept	−6.46	0.88	<0.001	−14.58	2.02	<0.001
Duration of diabetes (years)	0.23	0.03	<0.001	0.44	0.03	<0.001
Glycosylated hemoglobin level	0.13	0.03	<0.001	0.15	0.03	<0.001
Diastolic blood pressure	0.03	0.01	<0.001	0.03	0.01	<0.001
Body mass index	0.16	0.17	0.35	4.19	1.03	<0.001
(Duration of diabetes)^2	—	—	—	−0.007	0.0007	<0.001
(Body mass index)^2	—	—	—	−0.55	0.14	<0.001
Goodness-of-Fit						
	Statistic	d.f.	p-value	Statistic	d.f.	p-value
Q	62.99	31	<0.001	32.96	31	0.37
Q_R	59.48	31	<0.001	38.33	31	0.17

number of individuals in each covariate region may be small. The large number of additional parameters in the alternative hypothesis and the accompanying small frequencies in partitioned regions will affect the asymptotic chi-square distribution of the goodness-of-fit statistics. We suggest that data analysts adhere to the following recommendations when utilizing the proposed goodness-of-fit tests. First, if the data have a large number of time points (cluster size), one may consider partitioning the time variable as we did in partitioning the covariate space. Secondly, if the data set has a large number of covariates, one may consider an analysis of logit models with one covariate (or two covariates) together with the goodness-of-fit tests for variable selection (see the second illustrative example). If the goodness-of-fit tests are statistically significant, the analysts should further consider different forms of the covariates (e.g., adding quadratic or interaction terms) in the model. Using the results from this process, one can derive a possible final model with the selected functional forms of the covariates. Note that the number of covariates remaining in the model may still be large. To conduct goodness-of-fit tests, one may need to partition the space formed by this large number of covariates. To ensure the asymptotic property of the goodness-of-fit tests, the partitioned regions should contain sufficient observations. For covariate partitioning, we recommend that analysts apply the same suggested guidelines as when using weighted least squares (Koch et al., 1977; Forthofer and Lehnen, 1981): (a) Each partitioned region should contain at least 10 individuals (clusters); (b) Only 25% of the regions should have less than 25 individuals (clusters); (c) The frequency of $Y = 1$ or $Y = 0$ in each region should be greater than zero.

In the illustrative example from the Wisconsin Diabetic Retinopathy Study, the number of individuals in each of the 16 regions ranged from 16 to 76, with only 1 region having less than 25 individuals. None of the regions contains zero frequencies of $Y = 1$ or $Y = 0$. In practice, if the above criteria in region partitioning are not met, one may consider combining regions, while recognizing that the statistical power may decrease.

The main drawback of the proposed goodness-of-fit tests is that the test statistics and the corresponding degrees of freedom depend on the choice of covariate partitioning. In practice, investigators often know the usual ranges of the collected covariates. To avoid data dredging, one may decide the cut-points of continuous covariates before looking at the data. Nevertheless, the proposed tests provide a global summary of how well a model fits the data. When the proposed goodness-of-fit tests are statistically significant, it raises a red flag to the analyst to further examine the data and to perform additional explorations. As in any data analysis, the summary measure of fit is not enough for assessing the fit of the model. One should also examine the residuals of each of the observations for further confirmation.

Alternative goodness-of-fit tests for GEE modeling with binary responses may be developed by using some nonparametric smoothing methods. For example, le Cessie and Van Houwelingen (1991) proposed a goodness-of-fit test for logistic regression models with binary responses using sums of smoothed residuals. They used a kernel smoother based on distances in the covariate space. Furthermore, they extended this method to a score test for generalized linear models for assessing model fit (le Cessie and Van Houwelingen, 1995). They considered adding a random effect term to the regression model instead of adding additional parameters based on partitioned regions in the alternative hypothesis and proposed a goodness-of-fit test by testing that the variance of this extra random component equals zero. Hence, their test avoided the explosion of a large number of additional parameters and dependence on covariate partitioning. However, applying their test to models for correlated binary responses requires further research and is beyond the scope of this paper.

Acknowledgements

This research was supported in part by the University Research Committee of Emory University (HXB) and by National Institutes of Health grant T32-AI07442 (JMW). We would like to thank Jamie Grady and Tom Stiger for critical reading of the manuscript.

Résumé

L'analyse des données en mesures répétées est souvent accomplie en utilisant la méthodologie des équations d'estimation généralisées (GEE) (Liang et Zeger, 1986, *Biometrika* **73,** 13–22). Bien qu'il existe des méthodes pour évaluer l'adéquation de modèles ajustés sur des données non corrélées avec des méthodes de vraisemblance, il n'est pas approprié d'utiliser de telles méthodes pour des modèles ajustés avec la méthodologie des GEE. Nous proposons des tests d'ajustement basés sur un modèle et robustes (corrigés empiriquement) pour la modélisation de GEE avec des réponses binaires, reposant sur la partition de l'espace des covariables en des régions distinctes, et en formant des statistiques du score qui suivent asymptotiquement une distribution du Chi-2 avec le nombre de degrés de liberté approprié. La distribution sous l'hypothèse nulle et la puissance statistique des tests d'ajustement proposés sont étudiés en utilisant des données simulées. Les tests d'ajustement proposés sont illustrés par deux exemples utilisant des données provenant d'études cliniques.

REFERENCES

Agresti, A. (1990). *Categorical Data Analysis.* New York: Wiley.

Emrich, L. J. and 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.

Forthofer, R. N. and Lehnen, R. G. (1981). *Public Program Analysis: A New Categorical Data Approach.* Belmont, California: Lifetime Learning Publication.

Hosmer, D. W. and Lemeshow, S. L. (1989). *Applied Logistic Regression.* New York: Wiley.

Hosmer, D. W., Hosmer, T., Lemeshow, S. L., and le Cessie, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine* **16,** 965–980.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D., and 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. *Archives of Ophthalmology* **102,** 520–526.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H., and Lehnen, R. G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* **33,** 133–158.

Le Cessie, S. and van Houwelingen, J. C. (1991). A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics* **47,** 1267–1282.

Le Cessie, S. and van Houwelingen, J. C. (1995). Testing the fit of a regression model via score tests in random effects models. *Biometrics* **51,** 600–614.

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

Lipsitz, S. R. and Buoncristiani, J. F. (1994). A robust goodness-of-fit test statistic with application to ordinal regression models. *Statistics in Medicine* **13,** 143–152.

Tsiatis, A. A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika* **67,** 250–251.

Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42,** 121–130.

Received April 1996; *revised May and September* 1997; *accepted October* 1997.