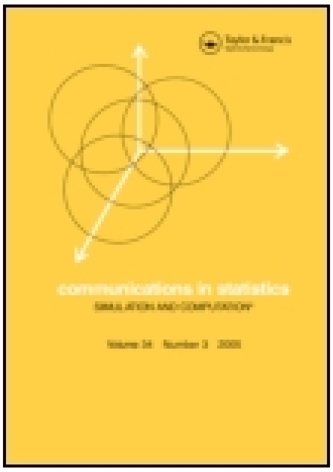

Communications in Statistics - Simulation and Computation
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lssp20

Goodness of Fit Tests for Logistic GEE Models: Simulation Results

Scott R. Evans Ph.D. [a c] & David W. Hosmer Jr., Ph.D. [b]

[a] Center for Biostatistics in AIDS Research , Department of Biostatistics , Harvard University , Boston, Massachusetts, USA

[b] Department of Biostatistics and Epidemiology , University of Massachusetts , Amherst, Massachusetts, USA

[c] Center for Biostatistics in AIDS Research , Department of Biostatistics , Harvard School of Public Health , FXB-513, 651 Huntington Ave., Boston, MA, 02115, USA
Published online: 20 Aug 2007.

COMMUNICATIONS IN STATISTICS
Simulation and Computation®
Vol. 33, No. 1, pp. 247–258, 2004

Goodness of Fit Tests for Logistic GEE Models: Simulation Results

Scott R. Evans, Ph.D.,[1,*] and David W. Hosmer Jr., Ph.D.[2]

[1]Center for Biostatistics in AIDS Research, Department of Biostatistics, Harvard University, Boston, Massachusetts, USA
[2]Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, Massachusetts, USA

ABSTRACT

Generalized Estimating Equations (GEE) have become a popular regression method for analyzing clustered binary data. Methods to assess the goodness of fit of the fitted models have recently been developed. However, published evaluations of these methods under various scenarios are limited. Research conducted for the ordinary logistic regression model provided the basis for a newly computed mean and variance for the Pearson statistic and the unweighted sums of squares statistic for the GEE case. A simulation study was conducted to evaluate the performance of these statistics under

*Correspondence: Scott R. Evans, FXB-513, Center for Biostatistics in AIDS Research, Department of Biostatistics, Harvard School of Public Health, 651 Huntington Ave., Boston, MA 02115, USA; Fax: +617-432-3163; E-mail: evans@sdac.harvard.edu.

DOI: 10.1081/SAC-120028443 0361-0918 (Print); 1532-4141 (Online)

various scenarios. The factors that were varied were the number of clusters, the number of observations within a cluster, the magnitude of the correlation, and the number and type of covariates included in the model. Overall, the Pearson and unweighted sums of squares statistics had a satisfactory performance of a type I error rate and are potentially effective in evaluating goodness of fit under certain conditions.

Key Words: Goodness of fit; GEE; Logistic regression.

1. INTRODUCTION

Clustered binary data may arise in practice in several ways. For example, repeated measurements on individual patients in a longitudinal study of the presence or absence of a disease often yield clusters of correlated observations. Observations taken on a particular subject will tend to be less variable than observations taken on different subjects. Thus observations taken on the same individual are considered to be dependent or clustered.

Clustered data results in a variance of response that is greater than that which would occur under independence. For binary data, this is often called extra binomial variation or overdispersion (Diggle et al., 1994). Failure to account for overdispersion may result in elevated type I errors as statistical tests will overstate the significance of differences in proportions (Pendergast et al., 1996). Well developed methods that assess model fit for the independent/non-clustered data setting, called here ordinary logistic regression (OLR) are not appropriate for clustered data due to the fact that residuals within the same cluster are correlated. The objective of this research is to evaluate summary statistics recently developed for logistic regression modeling with clustered binary data under various simulation scenarios.

2. THE GEE LOGISTIC REGRESSION MODEL

Liang and Zeger (1986) proposed GEE estimation to extend the OLR model for the analysis of clustered binary data. Their model describes how the marginal or average response across subjects changes with the covariates. Interest focuses on the relationship between the covariates and the probability of response while response correlation is treated as a nuisance parameter. This model is the natural analog of OLR and allows for the adjustment of subject (observational) and cluster level covariates. GEE is applicable for models that are unknown.

The logistic GEE model for the marginal or population averaged distribution of the binary outcome for observation j in cluster i, Y_{ij}, assumes that

$$\text{logit}[\pi(\mathbf{x}_{ij})] = \alpha + \boldsymbol{\beta}'\mathbf{x}_{ij} \tag{1}$$

where

$$\text{logit}(\pi(\mathbf{x}_{ij})) = \ln[\pi(\mathbf{x}_{ij})/\{1 - \pi(\mathbf{x}_{ij})\}],$$

for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, n_i$. In Eq. (1) the quantity $\pi(\mathbf{x}_{ij})$ is the marginal probability of response, $\mathbf{x}_{ij}$ is the p by 1 vector of covariates for observation j in cluster i, α is the population averaged intercept term, and $\boldsymbol{\beta}$ is the p by 1 vector of population averaged coefficients.

Parameters are then estimated by solving score-like equations called generalized estimating equations (Liang and Zeger, 1986). In order to solve the generalized estimating equations, a "working" correlation structure must be hypothesized for the within-cluster correlation. The GEE are of the form

$$\sum_{i=1}^{m} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{S}_i = 0$$

where $\mathbf{D}_i$ is the matrix of derivatives of $\pi(\mathbf{x}_{ij})$ with respect to model parameters, $\mathbf{S}_i$ is the vector of residuals where the individual components are $Y_{ij} - \pi(\mathbf{x}_{ij})$, the working covariance matrix is $\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}\mathbf{A}_i^{1/2}$, where $\mathbf{A}_i = \text{diag}(\text{var}(Y_{ij}))$ and $\mathbf{R}$ is the working correlation matrix (Liang and Zeger, 1986) which is usually unknown and must be estimated under the hypothesized form. Parameter estimates are consistent and asymptotically normal even if $\mathbf{R}$ is mispecified (Liang and Zeger, 1986).

3. GOODNESS OF FIT TESTS FOR OLR

The unweighted sums of squares and Pearson statistics are commonly used with to assess model fit in OLR. OLR may be viewed as a special case of GEE in which the size of every cluster is one. To simplify the notation for the OLR setting, we drop the second subscript and let the first denote the actual observation, $i = 1, 2, \ldots, n$ where $n = \sum_{i=1}^{I} n_i$. In this case the unweighted sums of squares and Pearson statistics are

$$\hat{S} = \sum_{i=1}^{n} (y_i - \hat{\pi}_i)^2, \quad \mathrm{X}^2 = \sum_{i=1}^{n} r(y_i, \hat{\pi}_i)^2$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

where

$$r(y_i, \hat{\pi}_i) = (y_i - \hat{\pi}_i)/(\hat{\pi}_i(1 - \hat{\pi}_i))^{1/2}$$

and $\hat{\pi}_i$ denotes the estimate of π_i. Moment estimators of these statistics have been developed for both of these statistics in the OLR setting.

le Cessie and van Houwelingen (1991) utilized first order Taylor series approximations to obtain an approximation for the estimated fitted values and the estimated residuals as a function of the true fitted values and the true residuals, respectively. In particular, they show that the vector of estimated fitted values may be expressed as a function of the vector of residuals $\mathbf{e} = \mathbf{y} - \boldsymbol{\pi}$ as $\hat{\boldsymbol{\pi}} \approx \boldsymbol{\pi} + \mathbf{H}_1\mathbf{e}$. The estimated residuals may be expressed as $\hat{\mathbf{e}} \approx (\mathbf{I} - \mathbf{H}_1)\mathbf{e}$ where $\mathbf{H}_1$ is a form of the hat matrix as follows

$$\mathbf{H}_1 = \mathbf{AX}(\mathbf{X}'\mathbf{AX})^{-1}\mathbf{X}'.$$

where $\mathbf{X}$ is an n by $(\mathrm{p} + 1)$ design matrix of the data.

Hosmer et al. (1997) using results from Oxius and Rojek (1992) obtained the moments of the Pearson statistic and the unweighted sums of squares statistic. They showed that each can be written as linear combinations of the true residuals. Under the assumption that the estimated probabilities and the estimated diagonal variance matrix are constant, their simulations showed that the standardized Pearson statistic, $Z_{\chi^2} = (\chi^2 - n)/\sqrt{\hat{\mathbf{a}}'(\mathbf{I} - \hat{\mathbf{H}}_1)\hat{\mathbf{A}}\hat{\mathbf{a}}}$ with $\hat{\mathbf{a}}' = (\mathbf{1} - 2\hat{\boldsymbol{\pi}})'\hat{\mathbf{A}}^{-1}$ has approximately a normal distribution with mean zero and a variance of one. Thus tests of fit can be performed by comparing Z_{χ^2} to percentiles of the $\mathrm{N}(0, 1)$ distribution. Similar results were obtained from simulations using a standardized unweighted sums of squares statistic, $Z_S = [\hat{S} - Trace(\hat{\mathbf{A}})]/\sqrt{\hat{\mathbf{b}}'(\mathbf{I} - \hat{\mathbf{H}}_1)\hat{\mathbf{A}}\hat{\mathbf{b}}}$ with $\hat{\mathbf{b}}' = (\mathbf{1} - 2\hat{\boldsymbol{\pi}})'$.

4. GOODNESS OF FIT TESTS FOR THE GEE LOGISTIC REGRESSION MODEL

Moment estimators of Z_{χ^2} and Z_S for logistic regression modeling for clustered binary data with estimation via GEE were developed independently by Evans (1998) and Pan (2002). Their moment estimates use a modified hat matrix defined as follows:

$$\mathbf{H}_2 = \mathbf{D}(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{V}^{-1}.$$

Following an approach similar to that of Hosmer et al. (1997), Evans (1998) and Pan (2002) approximate the moments of the Pearson statistic as follows:

$$E[\mathrm{X}^2 - n] \approx 0 \tag{2}$$

and

$$Var[\mathrm{X}^2 - n] \approx \mathbf{a}'(\mathbf{I} - \mathbf{H}_2)\mathbf{V}\mathbf{a}. \tag{3}$$

Similarly, it follows that the mean of the unweighted sums of squares statistic is approximately

$$E[\hat{S}] \approx Trace(\hat{\mathbf{A}}) \tag{4}$$

and the variance of the centered statistic is approximately

$$Var[\hat{S} - Trace(\hat{\mathbf{A}})] \approx \mathbf{b}'(\mathbf{I} - \mathbf{H}_2)\mathbf{V}\mathbf{b}. \tag{5}$$

These results suggest the following standardized statistics to test for goodness of fit following estimation by GEE:

$$Z_{\chi^2}{}^C = (\chi^2 - n)/\sqrt{\mathbf{a}'(\mathbf{I} - \mathbf{H}_2)\mathbf{V}\mathbf{a}} \tag{6}$$

and

$$Z_S^C = [S - Trace(\mathbf{A})]/\sqrt{\mathbf{b}'(\mathbf{I} - \mathbf{H}_2)\mathbf{V}\mathbf{b}} \tag{7}$$

and calculating significance levels using the N(0, 1) distribution.

5. SIMULATION STUDY

We performed a simulation study to evaluate the use of the approximations for the moments of the unweighted sums of squares and Pearson statistics developed by Evans (1998) and Pan (2002) under various simulation scenarios. The primary goal was assessment of the empirical type I error rates obtained via the standardized statistics in (6) and (7) since, in practice, the primary use of these statistics is to test model fit. A secondary goal was to examine the simulated means and variances of the statistics and compare them to the estimated means and variances in Eqs. (2)–(5).

The factors manipulated to determine the effects on the performance of the statistics were: (1) dimension (the number of clusters and the number of observations within a cluster), (2) covariates (the number and type of covariates included in the model), and (3) correlation. A factorial simulation study design was employed.

For each setting, clustered data were generated with an exchangeable correlation structure (i.e., all pairs of observations within a cluster are equally correlated) such that a mixed effects logistic model was the correct model. An example of the data generation procedure using SAS is provided in the Appendix. Correlation was controlled by the size of the coefficient of the random effect (generated as a standard normal variable). Data were generated such that the intercept was zero and the coefficients for all covariates in the model were equal to 0.8. This provided a reasonable set of values, which when combined with the different covariate distributions provided enough variation in the distribution of estimated probabilities. A model was then fitted using SAS macros assuming an exchangeable correlation structure to calculate the statistics of interest and the new estimates of their respective moments. Five hundred replications were performed for each combination of factors.

Statistical tests of the hypotheses that the standardized statistics are normal were conducted. No significant deviations from normality were observed for the unweighted sums of squares statistic. Some deviations were noted for the Pearson statistic, however, the distribution visually appeared as approximately normal (unimodal with approximately symmetric tails). Thus the significance level for each replication was obtained using the N(0,1) distribution.

Forty-five settings were examined utilizing five levels of dimension, three levels of covariates, and three levels of correlation as illustrated in Table 1. Table 2 displays the empirical rejection rates at alpha equal to 0.05 for each of the statistics at the various simulation settings.

Table 3 displays results summarizing simulated means and variances and estimates of the means and variances of the statistics from a few selected settings. If the estimated means of the Pearson and unweighted sums of squares statistics are correct, then the simulated means of these centered statistics (displayed in rows 1 and 4 of Table 3, respectively) should be approximately zero. If the estimated variances of the Pearson and unweighted sums of squares statistics are correct, then the variance estimates should approximate the simulated variances of the respective statistics (i.e., rows 2 and 3 should be nearly equivalent and rows 5 and 6 should be nearly equivalent).

Table 1. Simulation study factors and levels.

Factor A: Dimension	
A1:	200 clusters with 2 observations per cluster
A2:	100 clusters with 4 observations per cluster
A3:	80 clusters with 5 observations per cluster
A4:	50 clusters with 8 observations per cluster
A5:	20 clusters with 20 observations per cluster
Factor B: Covariates	
B1:	1 cluster level covariate with a standard normal distribution and 1 subject level covariate with a standard normal distribution
B2:	1 cluster level covariate only, with a standard normal distribution
B3:	1 subject level covariate only, with a standard normal distribution
Factor C: Correlation	
C1:	Low correlation ≈ 0.2
C2:	Medium correlation ≈ 0.5
C3:	High correlation ≈ 0.8

5.1. The Unweighted Sums of Squares Statistic

The unweighted sums of squares statistic performed particularly well for nearly all settings involving 50 or more clusters (see Table 2). It also performed well in settings involving 20 clusters and 20 observations per cluster when the model contained only a single covariate. In these cases the empirical alpha level was within sampling variability ($\pm 2\%$) of the target hypothesized alpha level, 0.05. The exceptions occurred in settings involving 20 clusters with 20 observations per cluster with two covariates. In these cases, the statistic rejected too often. Results were similar at the 0.10 and 0.01 alpha levels. These results suggest that the statistic is rejecting the correct percentage of the time for a given alpha under certain conditions.

In addition the simulated distribution of the "centered" statistic had a mean near zero for all settings that included 50 or more clusters (see Table 3). The variance estimate is not as close, however, this did not seem to have a significant adverse effect on the performance of the test statistic with respect to a type I error rate.

The unweighted sums of squares statistic is recommended for cases where the number of clusters is large (i.e., 50 or greater).

Table 2. Empirical rejection rates (%) based on 500 replications using alpha equal to 0.05 for each simulation setting. Approximate confidence intervals are obtained using ±2%.

Setting	$\hat{S}$	Pearson	Setting	$\hat{S}$	Pearson
A1/B1/C1	3.0	3.2	A4/B1/C1	5.8	5.8
A1/B1/C2	4.8	5.0	A4/B1/C2	5.2	5.8
A1/B1/C3	3.8	3.8	A4/B1/C3	3.0	7.0
A1/B2/C1	4.6	3.6	A4/B2/C1	4.6	4.2
A1/B2/C2	5.8	5.8	A4/B2/C2	4.4	3.6
A1/B2/C3	4.2	4.2	A4/B2/C3	5.2	4.0
A1/B3/C1	5.8	4.6	A4/B3/C1	3.0	5.6
A1/B3/C2	6.0	7.0	A4/B3/C2	3.4	5.0
A1/B3/C3	6.0	6.0	A4/B3/C3	6.0	6.4
A2/B1/C1	3.6	5.0	A5/B1/C1	7.6*	6.6
A2/B1/C2	7.4*	5.0	A5/B1/C2	9.4*	7.8*
A2/B1/C3	6.8	6.0	A5/B1/C3	13.7*	12.6*
A2/B2/C1	5.0	3.8	A5/B2/C1	3.2	3.0
A2/B2/C2	4.0	3.8	A5/B2/C2	6.0	6.2
A2/B2/C3	5.6	5.8	A5/B2/C3	4.2	4.4
A2/B3/C1	2.4*	5.2	A5/B3/C1	3.2	4.0
A2/B3/C2	4.8	4.6	A5/B3/C2	6.4	6.4
A2/B3/C3	5.6	4.6	A5/B3/C3	7.4*	6.8
A3/B1/C1	3.6	4.8			
A3/B1/C2	3.6	4.8			
A3/B1/C3	5.6	5.2			
A3/B2/C1	4.0	2.2*			
A3/B2/C2	6.2	5.2			
A3/B2/C3	5.2	4.0			
A3/B3/C1	3.2	5.6			
A3/B3/C2	5.0	6.4			
A3/B3/C3	5.0	4.0			

*Approximate confidence interval does not include targeted 5%.

5.2. The Pearson Statistic

The Pearson statistic utilizing the normal distribution performed particularly well for nearly all settings involving 50 or more clusters (see Table 2). In nearly all of these settings the empirical alpha level was within sampling variability of hypothesized value (i.e., 0.05). Similar to the unweighted sums of squares statistic, it did not perform well in settings involving only 20 clusters with two covariates. The statistic

Table 3. Simulation results from selected settings. All results based on 500 replications. Mean (standard deviation).

	Simulation setting			
	A1/B1/C1	A2/B2/C2	A4/B3/C2	A5/B1/C3
$Mean(\mathrm{X}^2 - n)$	−0.67 (7.69)	0.059 (1.35)	0.041 (5.74)	4.311 (32.3)
$Var(\mathrm{X}^2 - n)$	59.20	1.83	32.93	1041.44
$\hat{Var}(\mathrm{X}^2 - n)$	66.88 (60.79)	2.46 (6.43)	28.11 (14.14)	517.30 (1821)
$Mean(\hat{S} - Tr(\mathbf{A}))$	0.03 (0.465)	−0.003 (0.11)	0.027 (1.23)	0.666 (3.09)
$Var(\hat{S} - Tr(\mathbf{A}))$	0.22	0.012	1.513	9.523
$\hat{Var}(\hat{S} - Tr(\mathbf{A}))$	0.24 (0.138)	0.015 (0.028)	1.265 (0.515)	4.635 (7.40)

rejected too often for these cases when there was medium or high correlation. Results were similar at the 0.10 and 0.01 alpha levels.

The simulated sampling distribution of the "centered" Pearson statistic also had a mean near zero for settings with had at least 50 clusters (see Table 3). The variance estimate is not as close, however, this did not seem to have a significant adverse effect on the performance of the test statistic with respect to a type I error rate.

The Pearson statistic utilizing the normal distribution is recommended for use when the number of clusters is large (i.e., 50 or greater).

5.3. Summary

In summary, the performance of the unweighted sums of squares statistic and the Pearson statistic with the normal distribution are comparable. Each statistic appears to have an empirical type I error rate within sampling variability of the hypothesized alpha level for alpha equal to 0.10, 0.05, and 0.01 when the number of clusters is 50 or greater.

6. DISCUSSION

A simulation study was performed to evaluate use of the Pearson statistic and the unweighted sums of squares statistic under various scenarios for clustered binary data and estimation via GEE. Employing mean and variance estimates for these statistics discussed in Evans (1998) and Pan (2002), the size of the test (empirical type I error rate) for each test appears satisfactory under certain conditions. In particular, the standardized statistics are recommended for general use as long as the number of clusters is 50 or greater. This finding is not unexpected since the GEE model itself is most appropriate when the number of clusters is large (Liang and Zeger, 1986). An important limitation to note is that the unweighted sums of squares and Pearson statistics should only be used when at least one continuous covariate is included in the model as in the simulation study. The proposed tests may not perform well if only discrete covariates are included in the model.

Three major types of goodness of fit test statistics exist for OLR: statistics based upon partitioning of the covariate space, statistics using groups based on the ranked estimated probabilities, and statistics based on comparing observed versus predicted values (Hosmer and Lemeshow, 2000). Methodology for assessing goodness of fit in logistic GEE models is less well developed. Barnhart and Williamson (1998) have developed a statistic based on partitioning of the covariate space into distinct regions and forming score statistics. Horton et al. (1999) have developed a statistic using predicted deciles of risk. Evans (1998) has also investigated use of Hosmer–Lemeshow statistic and noted inadequate results of the statistic in several simulation scenarios. A disadvantage of these test statistics is that they depend upon subjective partitioning. Evans (1998) and Pan (2002) discuss statistics comparing observed versus predicted values (i.e., using residuals). These statistics can be easily applied in practice and are natural extensions of test statistics currently used in the OLR model. However, performance of these test statistics has only been evaluated in limited scenarios (see Pan, 2002).

We have only evaluated the type I error rate. The power of the OLR model versions of these statistics has been examined for the OLR model by Hosmer et al. (1997). Based upon superior power, the authors recommend use of the standardized unweighted sums of squares and Pearson statistics. Thus it is reasonable that as the logistic GEE model "approaches" the OLR model via small correlation, the unweighted sums of squares and Pearson statistics should have power similar to that reported by Hosmer et al. (1997) for some model departures (e.g., incorrect link function, inadequate content of the linear predictor, and a non-Bernoulli variance). Pan (2002) reports power results in a limited number of settings. These results suggest that power is adequate (>60%) when the number of clusters is 100 and 200.

We note that a summary measure of goodness of fit is only one aspect of thorough model development and evaluation. It is particularly important that fitted models are biologically supported. Proper assessment of fit should also include examination of individual components of summary measures such as fit to individual clusters and individual observations and regression diagnostics.

We believe that further research is needed in the following areas: (1) examination of the power of the tests under various scenarios, (2) use of the individual components of summary measures, and (3) evaluation of the robustness of these results to other correlation structures and assumptions concerning the correlation structure.

APPENDIX

The following is example SAS code used to generate the clustered binary data.

```
/***********DATA CREATION BEGINS***************/
data cluster;
  retain seed1 441225;
  retain seed2 690002;
  seed1=seed1+&k;
  seed2=seed2+&k;
do i=1 to 100;              ** 100 clusters;
zi=rannor(seed1);           ** a cluster level covariate;
ai=rannor(seed1);           ** a random effect used to generate
                               the correlation;
do j=1 to 4;                ** 4 observations per cluster;
x1ij=rannor(seed1);         ** a subject level covariate;
gij=1*ai+0.8*x1ij+0.8*zi;   ** the logit-correlation is controlled by
                               the coefficient of ai and the coefficients
                               of the covariates are 0.8;
pij=exp(gij)/(1+exp(gij));  ** the probability;
if (ranuni(seed2) le pij) then yij=1; else yij=0;    **dependent variable;
keep i j zi x1ij yij;
output;
end;
end;
run;
/*************DATA CREATION ENDS*****************/
```

ACKNOWLEDGMENTS

We would like to thank John Buonnaccorsi, Ph.D., Edward Stanek III, Ph.D., and the reviewer for their helpful comments. This work was supported in part by the National Institutes of Health Statistical and Data Management Center grant number 5 U01 AI38855 and the Neurologic AIDS Research Consortium grant number 2U01 NS0322228-08A1.

REFERENCES

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

Diggle, P. J., Liang, K., Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Evans, S. R. (1998). Goodness of Fit in Two Models for Clustered Binary Data. Ph.D. dissertation, University of Massachusetts.

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., Lemeshow, S. (2000). *Applied Logistic Regression*. 2nd ed. New York: Wiley.

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

le Cessie, S., van Houwelingen, J. C. (1991). Goodness of fit test for binary regression models, based on smoothing methods. *Biometrics* 47:1267–1282.

Liang, K., Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73(1):13–22.

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.

Pan, W. (2002). Goodness of fit tests for GEE with correlated binary data. *Scand. J. Statist.* 29(1):101–110.

Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M., Fisher, M. R. (1996). A survey of methods for analyzing clustered binary data. *Int. Statist. Rev.* 64(1):89–118.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016