STATISTICS IN MEDICINE
*Statist. Med.* **18**, 213–222 (1999)

# GOODNESS-OF-FIT FOR GEE: AN EXAMPLE WITH MENTAL HEALTH SERVICE UTILIZATION

NICHOLAS J. HORTON[1]*, JUDITH D. BEBCHUK[1], CHERYL L. JONES[1], STUART R. LIPSITZ[2], PAUL J. CATALANO[2], GWENDOLYN E. P. ZAHNER[3] AND GARRETT M. FITZMAURICE[1]

[1] *Department of Biostatistics, Harvard School of Public Health, Boston, MA, U.S.A.*
[2] *Dana Farber Cancer Institute, Boston and Department of Biostatistics, Harvard School of Public Health, U.S.A.*
[3] *Department of Psychiatry, Harvard Medical School, Boston, MA, U.S.A.*

## SUMMARY

Suppose we use generalized estimating equations to estimate a marginal regression model for repeated binary observations. There are no established summary statistics available for assessing the adequacy of the fitted model. In this paper we propose a goodness-of-fit test statistic which has an approximate chi-squared distribution when we have specified the model correctly. The proposed statistic can be viewed as an extension of the Hosmer and Lemeshow goodness-of-fit statistic for ordinary logistic regression to marginal regression models for repeated binary responses. We illustrate the methods using data from a study of mental health service utilization by children. The repeated responses are a set of binary measures of service use. We fit a marginal logistic regression model to the data using generalized estimating equations, and we apply the proposed goodness-of-fit statistic to assess the adequacy of the fitted model. Copyright © 1999 John Wiley & Sons, Ltd.

## INTRODUCTION

The use of generalized estimating equations to analyse repeated binary data has become increasingly common in the health sciences. However, few methods exist to assess the goodness-of-fit of the fitted marginal regression models. We propose a goodness-of-fit statistic that is an extension of the Hosmer and Lemeshow[1] statistic for ordinary logistic regression to marginal regression models for repeated binary observations. We illustrate the methods using data from a study examining mental health service utilization in children. A goal of this study is to explore the relationship between the repeated outcomes – binary measures of service utilization in one of

* Correspondence to: Nicholas J. Horton, Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A. E-mail: horton@hsph.harvard.edu

Contract/grant sponsor: NCI
Contract/grant numbers: CA-55576, CA-70101-01
Contract/grant sponsor: NIMH
Contract/grant numbers: T32-MH17119, U01-MH 51465, R01-MH54693
Contract/grant sponsor: NIEHS
Contract/grant number: ES-06900
Contract/grant sponsor: NIH
Contract/grant number: F31-GM 17274

CCC 0277–6715/99/020213–10$17.50

*Received October 1997*
*Accepted March 1998*

three settings (general health settings, school based settings, or mental health settings) – and covariates such as age, gender, family characteristics, and behavioural problems.

The models we consider are marginal models, which relate the expected value of each of the repeated binary responses separately to the covariates, via some appropriate link function (for example, logistic). To estimate the parameters of the marginal model, various authors have developed moment-based generalized estimating equations (GEEs),[2,3] which require specification of the form of the first two moments (the success probabilities and correlations) of the vector of binary responses for each individual. No established global measures of the goodness-of-fit of these marginal models are presently available. Our aims in this paper are to propose a goodness-of-fit statistic for marginal models and to illustrate its application.

For studies in which there are no repeated measures (that is, each subject has a single binary response), there have been several methods proposed for assessing the goodness-of-fit of logistic regression models. These are based on the notion of partitioning the subjects into groups or regions (Tsiatis[4] and Hosmer and Lemeshow[1]). We calculate a goodness-of-fit statistic as a quadratic form in the observed minus predicted responses in these regions or partitions. Schoenfeld also suggested a class of goodness-of-fit tests for the proportional hazards regression model.[5] Recently Barnhart and Williamson[6] have proposed a goodness-of-fit test that involves an extension of these methods to repeated outcomes.

With many continuous covariates and a single binary response, Hosmer and Lemeshow[1] propose partitioning subjects into groups or regions based on the percentiles of the predicted probabilities from the fitted logistic regression model. In this paper, we extend the Hosmer and Lemeshow method to repeated binary responses.

## GENERALIZED ESTIMATING EQUATIONS (GEE) FOR REPEATED BINARY OUTCOMES

In this section, we review the GEE methodology proposed by Liang and Zeger.[2,7] Assume that there are $T$ repeated measurements on $N$ subjects. However, either by design or due to missing data, the $i$th subject ($i = 1, \ldots, N$) has $T_i \leqslant T$ repeated observations. The binary random variable $Y_{it} = 1$ if the $i$th individual has response 1 (success) on measurement $t$, and $Y_{it} = 0$ otherwise. Each individual's data consist of a $(T_i \times 1)$ response vector $Y_i = [Y_{i1}, \ldots, Y_{iT_i}]'$, in addition to a $p \times 1$ covariate vector $\mathbf{x}_{it}$. The marginal density of $Y_{it}$ is Bernoulli

$$f(y_{it}|\mathbf{x}_{it}) = p_{it}^{y_{it}}(1 - p_{it})^{(1-y_{it})} \tag{1}$$

where we assume that

$$p_{it} = p_{it}(\boldsymbol{\beta}) = E(Y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}) = \mathrm{pr}(Y_{it} = 1|\mathbf{x}_{it}, \boldsymbol{\beta}) = \left(\frac{\mathrm{e}^{\beta_0 + \boldsymbol{\beta}_1'\mathbf{x}_{it}}}{1 + \mathrm{e}^{\beta_0 + \boldsymbol{\beta}_1'\mathbf{x}_{it}}}\right). \tag{2}$$

Although we assume a logit link function in (2), in principle, we can choose any suitable link function. Recall that our interest is in testing the adequacy of this logistic regression. Note that we can group the $p_{it}$'s together to form a vector $\mathbf{p}_i = \mathbf{p}_i(\boldsymbol{\beta})$ containing the marginal probabilities of success

$$\mathbf{p}_i(\boldsymbol{\beta}) = E[\mathbf{Y}_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT_i}, \boldsymbol{\beta}] = [p_{i1}, \ldots, p_{iT_i}]'.$$

 

We find the GEE estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, by solving the following estimating equations:

$$\mathbf{u}_\beta(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{N} \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} [\mathbf{Y}_i - \mathbf{p}_i(\hat{\boldsymbol{\beta}})] = 0 \tag{3}$$

where $\mathbf{D}_i = \partial \mathbf{p}_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$, and $\mathbf{V}_i$ is the $T_i \times T_i$ 'working' covariance matrix of $\mathbf{Y}_i$. We can specify this 'working' or approximate covariance matrix through a 'working' correlation matrix. In particular, we account for the correlation structure of each individual's vector of observations, $\mathbf{Y}_i$, by $\mathbf{R}_i(\boldsymbol{\alpha})$, a $T_i \times T_i$ 'working' correlation matrix, that is fully specified by an $s \times 1$ vector of unknown parameters $\boldsymbol{\alpha}$.

In (3), $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$, where $\mathbf{A}_i$ is a $T_i \times T_i$ diagonal matrix with $\text{var}(Y_{it}|\mathbf{x}_{it}) = p_{it}(1 - p_{it})$ as the $t$th diagonal entry. We obtain the estimate of $\boldsymbol{\beta}$ by replacing the unknown $\boldsymbol{\alpha}$ by any consistent estimate of it and solving for $\hat{\boldsymbol{\beta}}$ iteratively. Liang and Zeger[2] show that under mild regularity conditions, $\hat{\boldsymbol{\beta}}$ is a consistant estimator of $\boldsymbol{\beta}$.

## GOF FOR LOGISTIC REGRESSION

Before discussing our method for assessing the fit of the marginal model in (2), we briefly review the Hosmer–Lemeshow methodology for checking the goodness-of-fit for a logistic regression model (with no repeated measures). Since there are no repeated measures, for now, we drop the subscript $t$, and consider the fit of the following model:

$$\text{logit}(p_i) = \beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i. \tag{4}$$

Hosmer and Lemeshow suggest forming a set of $G$ groups, commonly 10. We form these groups based on the deciles of risk as determined by the estimated probabilities ($\hat{p}_i$) from model (4). In other words we use the lowest $n/10$ $\hat{p}_i$'s to form one category, we use the next highest $n/10$ $\hat{p}_i$'s to form a second category and so on. Thus we can create $(G - 1)$ new indicator variables to identify the decile of risk for each individual. One can show that the $\hat{p}_i$'s are a monotone transformation of an individual's covariates. For example, in the case where there is only one continuous covariate of interest, equivalent deciles of risk result whether they are based on the covariate of interest or the estimated $\hat{p}_i$'s. To determine if the model is a good fit, we add the additional $(G - 1)$ covariates to model (4). If we find these additional covariates significant, then there is evidence for lack of fit. Typically in the ordinary logistic regression setting, we use the Wald, likelihood ratio or score test statistic.

## GOF FOR REPEATED BINARY OUTCOMES

Consider the case where we have $T$ repeated measurements. Suppose we want to determine whether the (marginal) model

$$p_{it} = \left( \frac{e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_{it}}}{1 + e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_{it}}} \right).$$

is a good fit. One approach is to fit a broader model with interactions and/or polynomial and higher-order terms and to test whether the additional terms are significant. If they are not significant, then we may judge the model as having a good fit. Alternatively, we can obtain a 'global goodness-of-fit' statistic by extending the Hosmer–Lemeshow method. Following the

 

suggestion of Hosmer and Lemeshow for ordinary logistic regression, we propose forming $G$ (usually 10) groups based on combinations of the covariates $\mathbf{x}_{it}$'s in the logistic regression model, and testing to see if the additional regression coefficients for the $G-1$ indicator variables differ significantly from zero.

However, we need a rule for forming the groups based on combinations of the covariates $\mathbf{x}_{it}$. If all covariates are discrete, we can form a different group for each level in the cross-classification of covariates, but with many discrete covariates, there are too many groups. With many discrete and/or continuous covariates we suggest forming groups based on deciles of risk, as suggested by Hosmer and Lemeshow. That is, we form groups based on:

$$\tilde{p}_{it} = \left( \frac{e^{\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1' \mathbf{x}_{it}}}{1 + e^{\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1' \mathbf{x}_{it}}} \right).$$

Note, each subject has $T_i$ separate estimates of risk ($\tilde{p}_{it}$'s), and there are $\sum_{i=1}^{N} T_i$ observations in total. We suggest forming 10 groups of approximately equal size in the following manner:

1. The first group contains the $\sum_{i=1}^{N} T_i/10$, $(y_{it}, \mathbf{x}_{it})$'s with the smallest values of $\tilde{p}_{it}$.
2. The second group contains the $\sum_{i=1}^{N} T_i/10$, $(y_{it}, \mathbf{x}_{it})$'s with the next smallest values of $\tilde{p}_{it}$,

   ⋮

   and so on.

10. The last group contains the $\sum_{i=1}^{N} T_i/10$, $(y_{it}, \mathbf{x}_{it})$'s with the largest values of $\tilde{p}_{it}$.

Because some subjects may have identical covariate values, there can be ties in predicted risk, and so the number of subjects in each decile of risk may differ.

In general, we could form $G$ groups, with approximately $\sum_{i=1}^{N} T_i/G$ observations in each group. Since subject $i$ can have different $\tilde{p}_{it}$'s for each of the $T_i$ observations, a subject's group membership, $g$, can change for different $t$, $g = 1, \ldots, (G-1)$. That is, we can consider the group variable as a time-varying covariate. Observations in the same group have similar $\tilde{p}_{it}$'s and thus similar predicted 'risks'.

Suppose we define the $(G-1)$ group indicators

$$I_{itg} = \begin{cases} 1 \text{ if } \tilde{p}_{it} \text{ is in group } g \\ 0 \text{ if otherwise,} \end{cases} \qquad g = 1, \ldots, G-1$$

where the groups are based on 'percentiles of risk'. Then, to test the goodness-of-fit of the model given by (2), we consider the alternative model

$$\text{logit}(p_{it}) = \beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_{it} + \gamma_1 I_{it1} + \ldots + \gamma_{G-1} I_{it,G-1}.$$

Effectively, we are forming an 'alternative' model used to test the fit of the given model. Even though $I_{itg}$ is based on the random quantities $\tilde{p}_{it}$, Moore and Spruill[8] showed that, asymptotically, we can treat the partition as though based on the true $p_{it}$. Thus, we can treat $I_{itg}$ as a 'fixed' covariate. Results of limited simulations (not reported here, but available from the authors) suggest this assumption holds for moderately sized samples.

In general, if the model given by (4) is appropriate, then $\gamma_1 = \ldots = \gamma_{G-1} = 0$. A test of the fit of the model is equivalent to a test of:

$$H_0 : \gamma_1 = \cdots = \gamma_{G-1} = 0$$

 

which we can conduct using the GEE score or Wald statistic. Both statistics have asymptotic chi-squared distributions with $(G-1)$ degrees of freedom under the null hypothesis (that is, assuming the proposed model fits the data).

We propose the use of the score statistic since it only requires the estimate of $(\beta_0, \boldsymbol{\beta}_1)$ under the null, whereas the Wald statistic requires the estimate of $\boldsymbol{\gamma}_g$ under the alternative model. Also, the results of simulations suggest that the score statistic has better small sample properties. Finally, for a small percentage of simulation runs, the algorithm for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ in the alternative model (with $G-1$ 'extra' parameters) did not converge. When the algorithm does not converge, the Wald statistic is not available, but the score statistic is easily computable.

Next, we describe the score statistic in more detail. The quasi-score statistic for testing $H_0: \gamma_1 = \ldots = \gamma_{G-1} = 0$ within the GEE framework is of the following form. Letting $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_{G-1}]$, the score vector under the alternative ($H_1: \boldsymbol{\gamma} \neq \mathbf{0}$) is

$$\mathbf{u}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \begin{bmatrix} \mathbf{u}_1(\boldsymbol{\beta}, \boldsymbol{\gamma}) \\ \mathbf{u}_2(\boldsymbol{\beta}, \boldsymbol{\gamma}) \end{bmatrix} = \sum_{i=1}^{N} \begin{bmatrix} \mathbf{D}_{1i}' \mathbf{V}_i^{-1} [\mathbf{y}_i - \mathbf{p}_i(\boldsymbol{\beta}, \boldsymbol{\gamma})] \\ \mathbf{D}_{2i}' \mathbf{V}_i^{-1} [\mathbf{y}_i - \mathbf{p}_i(\boldsymbol{\beta}, \boldsymbol{\gamma})] \end{bmatrix} \tag{5}$$

where

$$\mathbf{D}_{1i} = \frac{\partial [\mathbf{p}_i(\boldsymbol{\beta}, \boldsymbol{\gamma})]}{\partial \boldsymbol{\beta}}, \quad \mathbf{D}_{2i} = \frac{\partial [\mathbf{p}_i(\boldsymbol{\beta}, \boldsymbol{\gamma})]}{\partial \boldsymbol{\gamma}}$$

and $\tilde{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$ under $H_0: \boldsymbol{\gamma} = 0$. Note that in the logistic regression setting, the first row of $D_{1i}$ is given by

$$\frac{\exp(\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_{it} + \boldsymbol{\gamma}' \mathbf{I}_{i1})}{(1 + \exp(\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_{it} + \boldsymbol{\gamma}' \mathbf{I}_{i1}))^2} (1, x_{i11}, x_{i12}, \ldots, x_{i1p})$$

and the first row of $\mathbf{D}_{2i}$ is given by

$$\frac{\exp(\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_{it} + \boldsymbol{\gamma}' \mathbf{I}_{i1})}{(1 + \exp(\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_{it} + \boldsymbol{\gamma}' \mathbf{I}_{i1}))^2} (I_{i11}, I_{i12}, \ldots, I_{i1,G-1}).$$

In general, the score test (Rotnitzky and Jewell[9]) is

$$X^2 = \mathbf{u}(\tilde{\boldsymbol{\beta}}, 0)' \{\widehat{\text{var}}[\mathbf{u}(\tilde{\boldsymbol{\beta}}, 0)]\}^{-1} \mathbf{u}(\tilde{\boldsymbol{\beta}}, 0). \tag{6}$$

However, we are using $\tilde{\boldsymbol{\beta}}$ which we obtain by solving $\mathbf{u}_1(\tilde{\boldsymbol{\beta}}, 0) = 0$. Therefore, the GEE score statistic for testing $H_0: \boldsymbol{\gamma} = 0$ is actually based on the large sample distribution of $\mathbf{u}_2(\tilde{\boldsymbol{\beta}}, 0)$, is given by

$$X^2 = \mathbf{u}_2(\tilde{\boldsymbol{\beta}}, 0)' \{\widehat{\text{var}}[\mathbf{u}_2(\tilde{\boldsymbol{\beta}}, 0)]\}^{-1} \mathbf{u}_2(\tilde{\boldsymbol{\beta}}, 0) \tag{7}$$

which is distributed as $\chi^2_{G-1}$ under the null. A SAS[10] macro was used to calculate the numerical value of equation (7). The macro used to conduct goodness-of-fit used the independence working covariance structure for computational simplicity. The variance estimate used in (7) is somewhat different than the one proposed by Rotnitzky and Jewell,[9] and appears in the appendix.

A significant GEE score statistic indicates that the proposed model leaves a substantial amount of variability in the data not taken into account.

## EXAMPLE: MENTAL HEALTH SERVICE UTILIZATION

In this section we apply the proposed methods to data from a study of mental health utilization by children. The study design has been reported elsewhere, [11, 12] as has a substantive analysis of

 

the service utilization data.[13] Subjects included 2519 children, aged 6–11, who were part of two cross-sectional surveys conducted in eastern Connecticut in the late 1980s. A goal of these surveys was to study determinants of mental health service utilization.

Parents of the children completed survey questionnaires that solicited information on child and household characteristics. The primary outcomes were service use in three settings: general health, school, and mental health. For a given setting, service use was defined as a parental report that the child had ever seen a provider or been in a special programme for a behavioural problem. If the particular service was used, the outcome was coded 1, and coded 0 otherwise. Clearly these binary outcomes are correlated for a given child.

In this study it is of interest to relate service use in the three settings to both child and family characteristics. Covariates measured for the child included age (AGE: 0 = age 6 to 8, 1 = age 9 to 11), gender (GENDER: 0 = female, 1 = male), ethnicity (BLACK: 0 = non-black, 1 = black and HISPANIC: 0 = non-Hispanic, 1 = Hispanic), and religion (CATHOLIC: 0 = non-Catholic, 1 = Catholic). The health and well-being of the child was assessed with the following covariates: total score on the Child Behavioral Checklist (CBCL) dichotomized at the 1991 published normal threshold (PQTOT: 0 = no problems, 1 = above borderline/clinical threshold for problems); academic problems (ACADPROB: 0 = no academic problems, 1 = repeated a grade, advised to repeat grade); and health problems (HLTHPROB: 0 = no health problems, 1 = fair or poor health, chronic condition, or limitation in activity).

Maternal and family covariates included belonging to a single parent household (MOMSING: 0 = father figure present, 1 = no father figure present); family stressors (FAMSTRESS: 0 = no stressors in past year, 1 = one or more stressors reported in past year); maternal reports of distress (MOMSTRESS: 0 = no home stress or family dissatisfaction reported by mother, 1 = report of home stress or family dissatisfaction); and whether the parent felt that their child needed special help or treatment for a problem noted on the CBCL (PQNDHLP: 0 = no need reported, 1 = need reported).

In the logistic regression analysis for repeated binary measures we adjusted for setting (using indicators for SCHOOL and MENTAL, that is, we used general services as baseline) and the above covariates. We assumed an independence working correlation structure and we obtained valid standard errors using the so-called 'sandwich' variance estimator.[2] We obtained similar results when we fit a model with an unstructured working covariance structure. Table I lists the parameter estimates and standard errors for the initial model having only main effects. These results indicated that being an older student, male, non-minority, non-Catholic, having more than the threshold number of problems on the CBCL, health problems, academic problems, parental report of need, and family stressors were all predictive of service use.

Next we consider testing the goodness-of-fit of this model. The goodness-of-fit test yielded a $\chi^2$ value of 17·29 and a p-value of 0·044. Since the goodness-of-fit statistic is based on 10 groups, it is distributed as $\chi^2_9$. This result suggests that there is substantial evidence that this model does not provide an adequate fit to the service use data.

Next, we considered additions to this main effects model to provide a better fit to the data. Table II displays the results from a model that includes interactions between several of the covariates and setting of service use. We introduced these interactions because prior studies suggested that the effects of certain problems depend on the service setting.[13] For example, one might expect that school based service use is associated with academic problems. Several of these interactions were highly significant, indicating their importance in modelling service utilization.

 

Table I. Parameter estimates and SEs (robust) for model with main effects only, obtained using independence working covariance matrix

| Parameter | Estimate | SE | z | *p*-value |
|---|---|---|---|---|
| INTERCEPT | − 3·828 | 0·162 | − 23·70 | < 0·0001 |
| SCHOOL | 1·075 | 0·096 | 11·25 | < 0·0001 |
| MENTAL | − 0·285 | 0·108 | − 2·64 | 0·008 |
| GENDER | 0·196 | 0·093 | 2·10 | 0·036 |
| AGE | 0·340 | 0·091 | 3·74 | 0·0002 |
| PQTOT | 0·616 | 0·111 | 5·54 | < 0·0001 |
| BLACK | − 0·880 | 0·152 | − 5·80 | < 0·0001 |
| HISPANIC | − 0·598 | 0·205 | − 2·92 | 0·003 |
| CATHOLIC | − 0·299 | 0·094 | − 3·17 | 0·002 |
| MOMSING | 0·195 | 0·130 | 1·50 | 0·135 |
| HLTHPROB | 0·343 | 0·092 | 3·73 | 0·0002 |
| ACADPROB | 0·814 | 0·103 | 7·89 | < 0·0001 |
| PQNDHLP | 1·530 | 0·101 | 15·12 | < 0·0001 |
| FAMSTRESS | 0·187 | 0·102 | 1·83 | 0·067 |
| MOMSTRESS | 0·147 | 0·117 | 1·26 | 0·208 |

Goodness-of-fit statistic $\chi^2 = 17{\cdot}29$, $p = 0{\cdot}044$

Table II. Parameter estimates and SEs (robust) for final model including higher-order interactions

| Parameter | Estimate | SE | z | *p*-value |
|---|---|---|---|---|
| INTERCEPT | − 3·338 | 0·183 | − 18·29 | < 0·0001 |
| SCHOOL | 0·237 | 0·190 | 1·25 | 0·212 |
| MENTAL | − 0·858 | 0·241 | − 3·56 | 0·0004 |
| GENDER | 0·199 | 0·095 | 2·09 | 0·036 |
| AGE | 0·350 | 0·093 | 3·76 | 0·0002 |
| PQTOT | 0·636 | 0·114 | 5·58 | < 0·0001 |
| BLACK | − 0·909 | 0·157 | − 5·80 | < 0·0001 |
| HISPANIC | − 0·615 | 0·212 | − 2·90 | 0·004 |
| CATHOLIC | − 0·304 | 0·097 | − 3·14 | 0·002 |
| MOMSING | 0·193 | 0·134 | 1·45 | 0·148 |
| HLTHPROB | 0·354 | 0·094 | 3·76 | 0·0002 |
| ACADPROB | 0·111 | 0·160 | 0·69 | 0·487 |
| PQNDHLP | 1·580 | 0·104 | 15·21 | < 0·0001 |
| FAMSTRESS | 0·003 | 0·170 | 0·02 | 0·985 |
| MOMSTRESS | 0·150 | 0·120 | 1·25 | 0·210 |
| ACADPROB*SCHOOL | 1·380 | 0·187 | 7·36 | < 0·0001 |
| ACADPROB*MENTAL | 0·088 | 0·212 | 0·42 | 0·676 |
| FAMSTRESS*SCHOOL | 0·088 | 0·201 | 0·44 | 0·663 |
| FAMSTRESS*MENTAL | 0·706 | 0·257 | 2·75 | 0·006 |

Goodness-of-fit statistic $\chi^2 = 13{\cdot}10$, $p = 0{\cdot}158$

We observed that academic problems were strongly associated with higher rates of school based service use, while having one or more family stressors was associated with higher rates of mental health setting service use. The proposed goodness-of-fit test provided no evidence for lack of fit for this model (GOF statistic $\chi^2 = 13{\cdot}10$, $p = 0{\cdot}158$).

 

## DISCUSSION

We have proposed a method to assess goodness-of-fit for repeated binary outcomes using predicted deciles of risk. We have illustrated this method with a predictive model for mental health service utilization. Guided by our proposed goodness-of-fit test, we rejected the initial model that contained main effects only in favour of a model that included higher-order interactions. This final model was consistent with results seen in previous analyses by Zahner and Daskalakis.[13]

Although the proposed goodness-of-fit statistic has a simple interpretation, due to its global nature it may miss important deviations from the fit, and can only directly test covariates that are in the model. Recent work by Hosmer *et al.*[14] reviewed a series of goodness-of-fit tests in the logistic regression setting and found they had little power in small samples. We conjecture that the same issues may apply in our setting. For a specific example, consider a situation where there is a quadratic relationship between a continuous covariate and the outcome. A test of non-linearity of that covariate would be much more powerful than our approach, since our omnibus statistic may not have high power to detect a specific alternative. Instead, it has broad based power to detect an array of general alternatives. This lack of power requires that the data analyst look for lack of fit in other ways. One should not consider a non-significant goodness-of-fit test as definitive evidence that a model is a good fit. We see the main value of this type of statistic in situations where it indicates lack of fit and prompts the data analyst to further exploration and to find ways to improve the model. Substantive knowledge of the subject-matter area should guide this process.

## APPENDIX

In this appendix, we derive $\widehat{\text{var}}[\mathbf{u}_2(\tilde{\boldsymbol{\beta}}, 0)]$ used in equation (7). This estimate can be considered a robust estimate of the conditional variance of $\mathbf{u}_2(\boldsymbol{\beta}, \boldsymbol{\gamma})$ given $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$, $\boldsymbol{\gamma} = \mathbf{0}$. We compute the estimate of variance by first taking a Taylor series expansion of $\mathbf{u}_1(\tilde{\boldsymbol{\beta}}, 0)$ in (5) to arrive at an estimate of $(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$. This expansion is given by

$$\mathbf{u}_1(\tilde{\boldsymbol{\beta}}, 0) \doteq \mathbf{u}_1(\boldsymbol{\beta}, 0) + \left[\frac{\partial \mathbf{u}_1(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\beta}}\right](\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Setting this equal to 0 yields

$$(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \doteq -\left[\frac{\partial \mathbf{u}_1(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\beta}}\right]^{-1} \mathbf{u}_1(\boldsymbol{\beta}, 0). \tag{8}$$

Similarly, we calculated a Taylor series expansion of $\mathbf{u}_2(\tilde{\boldsymbol{\beta}}, 0)$, plugging in the value of $(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ from (8). This expansion is given by

$$\mathbf{u}_2(\tilde{\boldsymbol{\beta}}, 0) \doteq \mathbf{u}_2(\boldsymbol{\beta}, 0) + \left[\frac{\partial \mathbf{u}_2(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\beta}}\right](\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

After substituting in $(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ from (8), $\mathbf{u}_2(\tilde{\boldsymbol{\beta}}, 0)$ becomes

$$\mathbf{u}_2(\tilde{\boldsymbol{\beta}}, 0) = \left[-\left(\frac{\partial \mathbf{u}_2(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \mathbf{u}_1(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\beta}}\right)^{-1}, \mathbf{I}\right]\begin{bmatrix}\mathbf{u}_1(\boldsymbol{\beta}, 0) \\ \mathbf{u}_2(\boldsymbol{\beta}, 0)\end{bmatrix}$$

where $\mathbf{I}$ is a $(G-1)\times(G-1)$ identity matrix.

 

This implies that the variance of $\mathbf{u}_2(\tilde{\boldsymbol{\beta}}, 0)$ is given by

$$\mathrm{var}[\mathbf{u}_2(\tilde{\boldsymbol{\beta}}, 0)] = \left[ -E\left(\frac{\partial \mathbf{u}_2(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\beta}}\right)\left\{E\left(\frac{\partial \mathbf{u}_1(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\beta}}\right)\right\}^{-1}, \mathbf{I}\right] E\left\{\begin{bmatrix}\mathbf{u}_1(\boldsymbol{\beta}, 0)\\ \mathbf{u}_2(\boldsymbol{\beta}, 0)\end{bmatrix}\begin{bmatrix}\mathbf{u}_1(\boldsymbol{\beta}, 0)\\ \mathbf{u}_2(\boldsymbol{\beta}, 0)\end{bmatrix}'\right\}$$
$$\times \left[ -E\left(\frac{\partial \mathbf{u}_2(\beta, 0)}{\partial \boldsymbol{\beta}}\right)\left\{E\left(\frac{\partial \mathbf{u}_1(\beta, 0)}{\partial \boldsymbol{\beta}}\right)\right\}^{-1}, \mathbf{I}\right]'.$$

We obtain the variance estimate by substituting

$$E\left\{\left(\frac{\partial \mathbf{u}_2(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \boldsymbol{\beta}}\right)\right\}_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}=0} \quad \text{for} \quad E\left(\frac{\partial \mathbf{u}_2(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\beta}}\right)$$

and

$$E\left\{\left(\frac{\partial \mathbf{u}_1(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \boldsymbol{\beta}}\right)\right\}_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}, \boldsymbol{\gamma}=0} \quad \text{for} \quad E\left(\frac{\partial \mathbf{u}_1(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\beta}}\right)$$

and

$$N^{-1}\sum_{i=1}^{N}\begin{bmatrix}\mathbf{u}_{1i}(\tilde{\boldsymbol{\beta}}, 0)\\ \mathbf{u}_{2i}(\tilde{\boldsymbol{\beta}}, 0)\end{bmatrix}\begin{bmatrix}\mathbf{u}_{1i}(\tilde{\boldsymbol{\beta}}, 0)\\ \mathbf{u}_{2i}(\tilde{\boldsymbol{\beta}}, 0)\end{bmatrix}'$$

for

$$E\left\{\begin{bmatrix}\mathbf{u}_1(\boldsymbol{\beta}, 0)\\ \mathbf{u}_2(\boldsymbol{\beta}, 0)\end{bmatrix}\begin{bmatrix}\mathbf{u}_1(\boldsymbol{\beta}, 0)\\ \mathbf{u}_2(\boldsymbol{\beta}, 0)\end{bmatrix}'\right\}$$

where the subscript $i$ refers to the appropriate score vector from subject $i$.

We note that, under an independence working assumption, the elements of $\mathbf{u}_2(\tilde{\boldsymbol{\beta}}, 0)$ are the observed minus expected values in the groups. If we then use the variance under independence in (6) and (7), we obtain identically the Tsiatis logistic regression statistic with partitions based on the predicted probabilities.

ACKNOWLEDGEMENTS

We are grateful for the support provided by NCI grants CA-55576 and CA-70101-01, NIMH grants T32-MH17119, U01-MH51465 and RO1-MH54693, NIEHS grant ES-06900 and NIH grant F31-GM17274. We would also like to thank Kathleen Propert, James Ware, Constantine Daskalakis, the editor and two anonymous reviewers for their helpful comments.

## REFERENCES

1. Hosmer, D. W. and Lemeshow, S. 'Goodness of fit tests for the multiple logistic regression model', *Communications in Statistics, Part A-Theory and Methods*, **9**, 1043–1069 (1980).
2. Liang, K-Y. and Zeger, S. L. 'Longitudinal data analysis using generalized linear models', *Biometrika*, **73**, 13–22 (1986).
3. Prentice, R. L. 'Correlated binary regression with covariates specific to each binary observation', *Biometrics*, **44**, 1033–1048 (1988).
4. Tsiatis, A. A. 'A note on a goodness-of-fit test for the logistic regression model', *Biometrika*, **67**, 250–251 (1980).
5. Schoenfeld, D. 'Chi-squared goodness of fit tests for the proportional hazards regression model', *Biometrika*, **67**(1), 145–153 (1980).
6. Barnhart, H. X. and Williamson, J. 'Goodness-of-fit tests for GEE modeling', *Biometrics*, **54**(2), 720–729 (1998).

 

7. Zeger, S. L. and Liang, K-Y. 'Longitudinal data analysis for discrete and continuous outcomes', *Biometrics*, **42**, 121–130 (1986).
8. Moore, D. S. and Spruill, M. C. 'Unified large-sample theory of general chi-squared statistics for tests of fit', *Annals of Statistics*, **3**, 599–616 (1975).
9. Rotnitzky, A. and Jewell, N. P. 'Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data', *Biometrika*, **77**, 485–497 (1990).
10. SAS/IML Software: Usage and Reference, Version 6, First Edition. SAS Institute Inc., Cary, NC, 1989.
11. Zahner, G. E. P., Pawelkiewicz, W., DeFrancesco, J. J. and Adnopoz, J. 'Children's mental health service needs and utilization patterns in an urban community', Journal of the American Academy of Child Adolescent Psychiatry, **31**, 951–960 (1992).
12. Zahner, G. E. P., Jacobs, J. H., Freeman, D. H. and Trainor, K. 'Rural-urban child psychopathology in a northeastern U.S. state: 1986–1989', *Journal of the American Academy of Child Adolescent Psychiatry*, **32**, 378–387 (1993).
13. Zahner, G. E. P. and Daskalakis, C. 'Factors associated with mental health, general health and school-based service use for child psychopathology', *American Journal of Public Health*, **87**(9), 1440–1448 (1997).
14. Hosmer, D. W., Hosmer, T., LeCessie, S. and Lemeshow, S. 'A comparison of goodness-of-fit tests for the logistic regression model', *Statistics in Medicine*, **16**, 965–980 (1997).

 