ELSEVIER

Computer Methods and Programs in Biomedicine 49 (1996) 199–210

computer methods and programs in biomedicine

A SAS macro for stepwise correlated binary regression

Isaac F. Nuamah[a*1], Yinsheng Qu[b], Saeid B. Amini[c]

[a]*University of Pennsylvania Cancer Center, 528 Blockley Hall, Philadelphia, PA 19104-6021, USA*
[b]*Department of Biostatistics and Epidemiology, Cleveland Clinic Foundation, Cleveland, OH 44106, USA*
[c]*Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106, USA*

Received 11 August 1995; accepted 9 January 1996

Abstract

Several regression methods have been proposed for the analysis of correlated binary data, but none deals with the selection of covariates when there exist a large number of potentially relevant covariates. We present a SAS macro based on a stepwise selection procedure for the analysis of correlated binary data. Using regression methods based on generalized estimating equations originally proposed by Liang and Zeger [1] and extended by Prentice [2], we describe a score test for forward selection, a Wald's test for backward elimination, and a test for model adequacy based on generalized scores. The methodology and the accompanying computer macro program written in SAS IML are illustrated with data from a prospective study of functional decline in the activities of daily living in a group of elderly patients.

Keywords: Correlated binary data; Generalized estimating equations; Longitudinal methods; Stepwise regression; Variable selection

1. Introduction

Regression analysis of correlated data arising from longitudinal designs have become increasingly popular in the literature. In most of these instances, the purpose is to describe a regression model for a function of the mean of the response variable. In this type of data, the response variables consist of either clustered data or repeated observations of some response of interest. These responses are usually correlated and the dependence among the responses needs to be accounted for in order to make correct inferences. For correlated binary data, the generalized estimating equations (GEE) approach proposed by Liang and Zeger [1] has been used as the basis for regression modeling. In this paper, we present a SAS macro, called STEPGEE, whose main focus is on the selection of covariates for these models when there exist a large number of potentially relevant covariates. For independent binary outcomes, its counterpart is the stepwise logistic procedure. This stepwise covariate selection procedure for correlated binary regression is based on Prentice's [2]

* Corresponding author.

[1] This work was done while the first author was at Case Western Reserve University, Cleveland, Ohio, USA

0169-2607/96/$15.00
PII S0169-2607(96)01718-X

extension of the original generalized estimating equations approach. We consider marginal models in which the marginal expectation of the binary responses is related to a set of potentially useful covariates by the logit link function.

In section 2, the model and estimation based on Prentice's [2] extension of Liang and Zeger's generalized estimating equations (Liang and Zeger [1], Zeger and Liang [3]) will briefly be discussed. In section 3, a stepwise covariate selection procedure will be presented using efficient score tests for forward selection and a modified Wald's test for backward elimination. A test for model adequacy using a generalized score test is also provided. In section 4, we will describe the SAS macro, STEPGEE, discussing the data input, user-supplied parameters, and its output. In section 5, we will illustrate this procedure with data on functional decline in a group of elderly subjects. We shall not concentrate on any of the potential problems associated with stepwise selection in general, and we hope that any such concerns will not distract from the main issue at hand, one of variable selection for correlated binary data.

The selection strategy implemented in STEPGEE is similar to the one used in PROC LOGIST [4] in that it is limited to only binary responses. We know of no other software that can implement a selection strategy using GEE, but we know of at least three other programs that use the GEE methodology for estimation of regression coefficients. Karim and Zeger [5] have a SAS macro that analyses longitudinal data using the GEE methodology, but cannot handle missing data. By appearance, our program is similar to the program of Karim and Zeger, since they are both SAS macros and require the SAS system and its add-on interactive matrix language (IML) [6] for implementation. Two other programs written in different computer languages are available that are modelled after the Karim and Zeger original program. The first one by Carey (available by sending a request for GEE to STATLIB.CMU.EDU) is a C version that is designed to be used as an S function. The other one, RMGEE by Davis [7], is a FORTRAN version that has the additional capability of handling missing data.

2. Regression methodology

Liang and Zeger successfully applied the approach of generalized estimating equations (GEE) to longitudinal data whose marginal means are related to covariates through generalized linear models [1,3,8,9]. Prentice [2] extended the Liang and Zeger approach to allow for joint inference on the regression parameters and pairwise correlations among the observations. In this paper, we consider the approach taken by Lipsitz, Laird and Harrington [10], in which the odds ratio is used to model the pairwise correlations, since it seems more appropriate for binary data. In implementing our version of GEE, we kept these objectives in mind: (a) that one is interested in relating the binary response, **Y**, to a set of covariates, **X**; (b) that one is interested in characterizing the degree of association between pairs of outcomes using the odds ratio; and (c) one is interested in selecting from a given set of covariates a reduced set of significant covariates. The second objective might even include describing the dependence of the association parameters on the covariates. However, in implementing STEPGEE this objective is not considered, that is, it has been assumed that the odds ratio between any pair of outcomes do not depend on any of the covariates.

Let us assume that the i-th individual is observed a total of n_i ($i = 1,2,...,N$) times, where the maximum is T times, i.e. $n_i \leq T$. The responses are binary, that is, the binary random variable $y_{it} = 1$, if subject i has response 1, the response of interest, e.g. success at time t, and 0 otherwise. Each individual has a $px1$ covariate vector $\mathbf{x}_{it}$ measured at time t_i, which can be both time-stationary or time-varying. Let $\mathbf{x}_{it} = (x_{it_1}, x_{it_2},.....,x_{it_p})$ and μ_{it} be, respectively, the vector of covariates and mean corresponding to y_{it}, the t-th outcome for individual i. Let $\mathbf{y}_i = (y_{i1}, y_{i2},....,y_{in_i})$ and let $\mathbf{X}_i$ be the $n_i \times p$ matrix with rows equal to $\mathbf{x}_{it}$, i.e. $\mathbf{X}_i = (x_{i1}, x_{i2},....,x_{in_i})$. The marginal distribution of Y_{it} is binary and can be represented as

$$f(y_{it}|\mathbf{x}_{it}) = \exp[y_{it}\theta_{it} - log\{1 + \exp(\theta_{it})\}]$$

where

$$\theta_{it} = \log\{\mu_{it}/(1 - \mu_{it})\} = x_{it}{}^T\beta$$

and

$$\mu_{it} = \mu_{it}(\beta) = E(Y_{it}) = Pr(Y_{it} = 1|x_{it},\beta)$$

is the probability of success at time t; and β is a $p \times 1$ vector of parameters. Here the link function is the logit, a natural choice for binary data. The marginal probabilities can also be put together to form a vector $\mu_i(\beta) = E(Y_i) = (\mu_{i1}, \mu_{i2},\ldots,\mu_{in_i})^T$. To estimate β, Liang and Zeger [1] and Prentice [2] consider generalized estimating equations of the form

$$U_1(\beta) = \sum_{i=1}^{N} D_i{}^T V_i{}^{-1}(y_i - \mu_i(\beta)) = 0 \qquad (1)$$

where $D_i = \partial\mu_i(\beta)/\partial\beta$, and V_i is a 'working' covariance matrix of Y_i. This working covariance matrix has the form

$$V_i = A_i{}^{1/2} R_i(\alpha) A_i{}^{1/2} \qquad (2)$$

where $A_i = \text{diag}\{\text{Var}(Y_{it})\} = \text{diag}\{\mu_{it}(1 - \mu_{it})\}$, and $R_i(\alpha)$ is a 'working' correlation matrix. Here, V_i is a function of both β and α, and is specified entirely by the marginal distributions, i.e. by β and $R_i(\alpha) = \text{corr}\ (Y_i)$. Lipsitz, Laird and Harrington [10] suggested using the odds ratio, represented here as γ, as an alternative to the correlation coefficient. The pairwise association is measured in terms of odds ratio between Y_{is} and Y_{it}, $i = 1,2,\ldots,n_i$, $s \neq t$, and defined as

$$\gamma_{ist} = \frac{Pr(Y_{is} = 1, Y_{it} = 1)Pr(Y_{is} = 0, Y_{it} = 0)}{Pr(Y_{is} = 1, Y_{it} = 0)Pr(Y_{is} = 0, Y_{it} = 1)} \qquad (3)$$

which for a 2×2 table is given as the cross-products ratio. Let the log-odds ratio be $\log(\gamma_{ist}) = h^T{}_{ist}\alpha$, where h_{ist} is a $q \times 1$ vector which specifies the form of the relation between the association parameters and the vector of covariates, and α is a $q \times 1$ vector of association parameters where q is arbitrary. Thus, γ is a $\{n_i(n_i - 1)/2\} \times 1$ vector of pairwise odds ratios, expressed as a function of a $q \times 1$ vector α. In some studies it may be reasonable to assume that the γs are identical, i.e assume a common pairwise odds ratio. In STEPGEE, the form of the correlation structure chosen determines how many parameters would be estimated. However, one cannot specify the form of the relation between the association parameters and the vector of covariates, that is, it is assumed that $\log(\gamma_{ist}) = \alpha$. For each cluster, define a $\{n_i(n_i - 1)/2\} \times 1$ vector such that $Z_i = (y_{i1}y_{i2}, y_{i1}y_{i3},\ldots, y_{in_i-1}y_{in_i})^T$. Let $\eta_i = E(Z_i)$, then for $s,t = 1,2,\ldots,n_i$, $s \neq t$, it has been shown by Mardia [11] that

$$E(Y_{is}Y_{it}) = P(Y_{is} = Y_{it} = 1) = \begin{cases} (1 - (\mu_{is} + \mu_{it})(1 - \gamma_{ist}) - [\{1 - (\mu_{is} + \mu_{it})(1 - \gamma_{ist})\}^{1/2} \\ \quad - 4(\gamma_{ist} - 1)\mu_{is}\mu_{it}]^{1/2})/2(\gamma_{ist} - 1), & \text{if } \gamma_{ist} \neq 1, \\ \mu_{is}\mu_{it}, & \text{if } \gamma_{ist} = 1 \end{cases} \qquad (4)$$

which depends on the parameter set $\delta = (\beta, \alpha)$ through $\mu_{is} = E(Y_{is})$, $\mu_{it} = E(Y_{it})$ and γ_{ist}, the pairwise odds ratio. Note that, if we let $\mu_{ist} = E(Y_{is}Y_{it})$ then

$$\text{corr}(y_{is}, y_{it}) = \frac{\mu_{ist} - \mu_{is}\mu_{it}}{[\mu_{is}(1 - \mu_{is})\mu_{it}(1 - \mu_{it})]^{1/2}} \qquad (5)$$

is the (s, t)-th element of $R_i(\alpha)$ in (1). Using arguments similar to Prentice [2], Lipsitz et al. [6] showed that a second set of estimating equations of the form (1) could be obtained as

$$U_2(\alpha) = \sum_{i=1}^{N} E_i{}^T W_i{}^{-1}(Z_i - \eta_i(\beta,\alpha)) = 0 \qquad (6)$$

where $E_i = \partial\eta_i(\beta, \alpha)/\partial\alpha$. Although η_i is a function of both α and β, we assume that β is fixed in η_i in (6), that is why in E_i, we take derivatives only with respect to α. The two sets of estimating equations (1) and (6) provide a unified approach to estimating $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_q)$. Following Prentice [2], the estimation of the parameter values of $\delta = (\beta, \alpha)$ can be obtained iteratively by using

$$\begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} \beta \\ \alpha \end{bmatrix} + \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{B} & \mathbf{C} \end{bmatrix} \begin{bmatrix} U_1(\beta) \\ U_2(\alpha) \end{bmatrix} \tag{7}$$

where $U_1(\beta)$ and $U_2(\alpha)$ have been defined in (1) and (6), and

$$A = \left(\sum_{i=1}^{N} D_i^T V_i^{-1} D_i \right)^{-1}$$

$$B = \left(\sum_{i=1}^{N} E_i^T W_i^{-1} E_i \right)^{-1} \left(\sum_{i=1}^{N} E_i^T W_i^{-1} \partial Z_i / \partial \beta^T \right) \times \left(\sum_{i=1}^{N} D_i^T V_i^{-1} D_i \right)^{-1}$$

$$C = \left(\sum_{i=1}^{N} E_i^T W_i^{-1} E_i \right)^{-1}$$

However, a computationally simpler version can be used in practice (Prentice [2]; Qu, Piedmonte and Williams [12]) by setting the matrix **B** to zero resulting in $p + q$ equations. STEPGEE prints out the values of the matrix **B** so that users can decide for themselves how good the resulting equations are in computing the parameter estimates. From our experience, this simpler version is very good and the values of **B** were always less than 9×10^{-3}. The resulting equations are

$$\hat{\beta} = \beta + \left(\sum_{i=1}^{N} D_i^T V_i^{-1} D_i \right)^{-1} \times \sum_{i=1}^{N} D_i^T V_i^{-1} (Y_i - \mu_i) \tag{8}$$

$$\alpha = \alpha + \left(\sum_{i=1}^{N} E_i^T W_i^{-1} E_i \right)^{-1} \times \sum_{i=1}^{N} E_i^T W_i^{-1} (Z_i - \eta_i) \tag{9}$$

Iterating between these two equations leads to consistent parameter and variance estimates. This form is similar to the alternating logistic regressions approach of Carey, Zeger and Diggle [13]. Let $\hat{\beta}$ be the GEE estimate obtained, and if the variance of Y is correctly specified, then the variance of $\hat{\beta}$ is estimated by

$$\hat{\text{var}}(\hat{\beta}) = \left(\sum_{i=1}^{N} \hat{D}_i^T \hat{V}_i^{-1} \hat{D}_i \right)^{-1} \tag{10}$$

The robust variance estimate, which is consistent even if the covariance structure of y is misspecified has been given as (Liang and Zeger [1], Zeger and Liang [3]):

$$\hat{\text{var}}(\hat{\beta}) = \left(\sum_{i=1}^{N} \hat{D}_i^T \hat{V}_i^{-1} \hat{D}_i \right)^{-1} \times \left(\sum_{i=1}^{N} \hat{D}_i^T \hat{V}_i^{-1} \hat{\text{cov}}(Y_i) \hat{V}_i^{-1} \hat{D}_i \right) \times \left(\sum_{i=1}^{N} \hat{D}_i^T \hat{V}_i^{-1} \hat{D}_i \right)^{-1} \tag{11}$$

3. Theoretical considerations

3.1. Asymptotic distributions

We shall now discuss some of properties of the score function $U(\beta)$, defined as

$$U(\beta) = \sum_{i=1}^{N} D_i^T V_i^{-1} (y_i - \mu_i(\beta)) \tag{12}$$

where V_i is the variance matrix of y_i. This score function has the following properties:

$$E(U) = 0 \tag{13}$$

$$E(-\partial U/\partial\beta) = \sum_{i=1}^{N} D_i{}^T V_i^{-1} D_i = J \qquad (14)$$

where J is the Fisher information, and the variance of U equals J. If the true variance of y_i is not known, and V_i is a 'working' covariance matrix of y_i, then the properties of (13) and (14) still hold, but the variance of the score U is now given as

$$\mathrm{Var}(U) = \sum_{i=1}^{N} D_i{}^T V_i^{-1} \mathrm{var}(y_i)\ V_i^{-1} D_i = \Sigma \qquad (15)$$

which is not equal to the information matrix J unless $V_i = \mathrm{var}(y_i)$ for all $i = 1,2,...,N$. Under mild regularity conditions (Rao [14], Liang and Zeger [1]), the distribution of the score function U is asymptotically multivariate Gaussian with mean zero and variance-covariance matrix Σ. Also, the distribution of $N^{1/2}\ (\hat{\beta} - \beta)$ is asymptotically multivariate Gaussian, with mean zero and non-singular variance-covariance matrix $NJ^{-1}\Sigma J^{-1}$. We note here that the method by which the asymptotic distribution of the GEE estimators are determined follows closely those of the maximum likelihood estimators, i.e. they involve the expansion of the score function by Taylor's theorem. The joint asymptotic distribution of $\delta = (\beta,\ \alpha)$ follows directly from the result above [2].

3.2. Hypothesis testing

From the preceding asymptotic probability distribution based on the generalized estimating equations, Wald and score-type tests can be constructed. To test the null hypothesis that the p-variate vector $\beta = \beta_O$, let $\hat{\beta}$ be the GEE estimate of β, then under the regularity conditions already alluded to, we can construct the Wald test statistic

$$(\hat{\beta} - \beta_o)^T \hat{V}_G{}^{-1} (\hat{\beta} - \beta_o) \qquad (16)$$

where $V_G = J^{-1}\Sigma J^{-1}$, which asymptotically under the null has a central chi-square distribution with p degrees of freedom. The variance term V_G is evaluated at the GEE estimates $\hat{\beta}$. If we define the null hypothesis as H_O:$\beta = 0$, then the test corresponds to testing that all p parameters are jointly zero. For the simple null hypothesis H_O: $\beta_k = 0$, for any k-th element from the p parameters, the Wald test corresponds to

$$T_W = (\hat{\beta}_k)^2 / V_{kk}(\hat{\alpha},\hat{\beta}) \sim \chi_1^2 \qquad (17)$$

where V_{kk} is the (k,k)-th element of the variance-covariance matrix V_G. An alternative method of testing the joint significance of all variates (βs) uses the score test statistic

$$U^T(\hat{\beta}_o)\Sigma^{-1}U(\hat{\beta}_o) \qquad (18)$$

where $\hat{\beta}_O$ is a vector of GEE estimates, and Σ is evaluated at $\hat{\beta}_O$. This statistic under the null hypothesis is distributed as a chi-square distribution with p degrees of freedom. The score statistic for testing the simple null hypothesis follows directly from (18) and we discuss this further in Section 4. In all these tests, the correlation parameter, α, is estimated and assumed fixed at its limiting value. Hence, whenever we write the score function, $U(\beta)$, the correlation parameter is suppressed.

3.3. Generalized score tests

Suppose the parameter vector β can be partitioned as $\beta = (\beta_1{}^T, \beta_2{}^T)^T$, where β_1 and β_2 are vectors of dimension $(p - r) \times 1$ and $r \times 1$, respectively. The score function U and its variance matrix Σ, as well as the information matrix J, can also be partitioned to $U(\beta) = (U_1{}^T(\beta),\ U_2{}^T(\beta))^T$,

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

and

$$J = \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix}$$

respectively. The null hypothesis H_O is $\beta_2 = \beta_{2O}$.

Let $\tilde{\beta} = (\tilde{\beta}^T_1, \beta^T_{2O})^T$ be the vector of GEE estimates under H_O. The score $\hat{U}(\hat{\beta}) = (\hat{U}_1^T(\hat{\beta}), \hat{U}_2^T(\hat{\beta}))^T = (0^T, \hat{U}_2^T(\hat{\beta}))^T$, where

$$U_2(\hat{\beta}) = \sum_i \frac{\partial \mu_i^T}{\partial \beta_2} \hat{V}_i^{-1}(y_i - \hat{\mu}_i) \tag{19}$$

The generalized score test for H_O is given by

$$T_U = \hat{U}_2{}^T(\hat{\beta}) \hat{V}_{U_2}{}^{-1} \hat{U}_2(\hat{\beta}) \tag{20}$$

where $\hat{V}_{U_2}$ is the variance matrix of the subvector $\hat{U}_2(\hat{\beta})$. $\hat{V}_{U_2}$ can be calculated from the submatrices of $\hat{\Sigma}_{11}$, $\hat{\Sigma}_{21}$, $\hat{\Sigma}_{22}$, $\hat{J}_{11}$, and $\hat{J}_{12}$ such that

$$\hat{V}_{U_2} = \hat{\Sigma}_{22} - \hat{J}_{21}\hat{J}_{11}^{-1}\hat{\Sigma}_{12} - \hat{\Sigma}_{21}\hat{J}_{11}^{-1}\hat{J}_{12} + \hat{J}_{21}\hat{J}_{11}^{-1}\hat{\Sigma}_{11}\hat{J}_{11}^{-1}\hat{J}_{12} \tag{21}$$

(Boos [15], Breslow [16]). Under the null hypothesis, T_U has a chi-square distribution with r degrees of freedom. For a model-based score function, since $\Sigma = J$, $\hat{V}_{U_2}$ reduces to

$$\hat{V}_{U_2} = \hat{J}_{22} - \hat{J}_{21}\hat{J}_{11}^{-1}\hat{J}_{12} \tag{22}$$

A test for model adequacy also based on the generalized score statistic is available in STEPGEE, by testing that the parameters not selected into the final model are jointly zero. Alternatively, one could use a Wald test to test for the significance of the selected parameters. Such a method is given in the program by Davis [7]. We note here that these assumptions have been made for the tests to be valid for the correlated binary regression model: (a) consistent variance estimates are obtained under the assumption that the weighted average of the estimated correlation matrices converge to a fixed matrix; (b) the GEE estimates of the parameters obtained under the reduced model and the correlation parameters are assumed to be fixed at their limiting value, when testing for the significance of the parameters not yet selected into the model; (c) the 'working' correlation matrix is assumed to be the true correlation structure during hypothesis testing; (d) for clusters of varying sizes, the correlation structure of a given size corresponds with the appropriate submatrix of the correlation structure of the clusters of larger size.

4. Variable selection

Variable selection strategies have received considerable attention in the literature. They include such strategies as forward selection, backward elimination, and subset selection. Although there is no 'best' strategy for variable selection, a reasonable and often-used approach is the stepwise selection, and this is the selection strategy employed in STEPGEE. Stepwise selection refers to the strategy of using forward selection followed by backward elimination. While this strategy is useful in identifying significant predictive variables from among a large set of other explanatory variables, cautious use of them leads to correct interpretation of the final model. Several authors have provided broad guidelines for use with any variable selection procedure. Among them are Krall [17], Harrell et al. [18], Miller [19] and Hauck and Miike [20].

To select 'significant' variables, we employ the score statistic defined in Section 3. Suppose that in the regression model, we have selected $p - 1$ variables and the GEE estimates of the regression coefficients are denoted by the vector β_1. Now a new covariate with a regression coefficient β_2 is added to the model. We want to test H_O: $\beta_2 = \beta_{2O} = 0$. Let $\hat{\beta} = (\hat{\beta}_1, 0)$, and calculate the generalized score statistic

$$T_U = \hat{U}_2{}^T(\hat{\beta})\ \hat{V}_{U_2}{}^{-1} \hat{U}_2(\hat{\beta})$$

using

$$\hat{V}_{U_2} = \hat{\Sigma}_{22} - \hat{J}_{21}\hat{J}_{11}^{-1}\hat{\Sigma}_{12} - \hat{\Sigma}_{21}\hat{J}_{11}^{-1}\hat{J}_{12} + \hat{J}_{21}\hat{J}_{11}^{-1}\hat{\Sigma}_{11}\ \hat{J}_{11}{}^{-1}\hat{J}_{12}$$

Under the null hypothesis, T_U follows the chi-square distribution with 1 degree of freedom. The new variable is selected when H_O is rejected. We have assumed that the significant covariates already selected into the model and the estimates of the correlation parameters are fixed at their

limiting values, and therefore do not affect the asymptotic distribution of T_U.

If more than one variable is outside the regression model, the forward selection procedure works by first calculating the T_U statistic for each of them and selecting the one with the largest T_U. If it is significant at a nominally set significant level, referred to as PIN in STEPGEE, then it is added to the model and GEE estimation procedure carried over again for the current number of parameters in the model.

After each forward selection step, tests for backward elimination begin. Suppose that at the i-th step, the model contains $i + 1$ variables (including the intercept term). We wish to test for each variable currently in the model (except the intercept term) which one should be eliminated. We test the hypothesis H_O: $\beta_k = 0$, using the Wald test which is given in Eq. (17). If any of T_W's is not significant at the chosen level of significance, POUT, the least significant one is dropped from the model, otherwise the forward selection step is repeated. We also opted to use the t-distribution for backward elimination instead of the normal distribution for the Wald ratio since an earlier simulation study on the small sample properties of correlated binary models seem to indicate that tests based on the t-distribution results in smaller type I error coverage probability [12]. The degrees of freedom associated with the t-distribution are (total number of clusters – total number of parameters). The number of parameters consist of the number of correlation parameters and the number of parameters selected into the model.

STEPGEE implements this stepwise procedure based on the forward selection followed and the backward elimination strategies just described. The forward addition is immediately followed by a test for backward elimination. If no additional variable can be added to the model at the preset PIN level, or the variable just added is the least significant one to be deleted, then the stepwise procedure stops, indicating no change in the model. The final model significance is tested by the generalized score test for model adequacy. The final decision regarding the suitability of such a model should be made considering all available a priori information regarding the variables and their effect on outcome. STEPGEE is only a guide in this general direction.

5. Program description

STEPGEE is a SAS macro for performing stepwise regression analysis when the outcome consists of repeated measurements of a binary response variable. Since the program was written using the IML procedure of SAS, it assumes that the user is familiar with a basic SAS DATA step setup and has access to the SAS system. An input data file has to be created using the PROC DATA step of SAS. The macro facility then allows the user to pass arguments to STEPGEE. The limitation on the size of input data depends on the version of SAS available at one's site. For example, the program yielded an out-of-memory error when run on the DOS version of SAS with our sample data set, but worked fine in Windows, UNIX and OS platform versions. The time taken to implement one run also depends on the number of input parameters and the speed of the computer's microprocessor. The program prints out a description of data, and some parameters that were passed to the program. It prints out at each selection or deletion, the parameters and their estimated coefficients, robust standard errors and a P-value based on a t-statistic. It also prints out the naive standard errors of the selected variables at each step. In addition, it prints out the parameters not currently selected into the model, together with their associated Wald's tests P-values. The working correlation matrix selected is also printed. These parameters are required in the macro step to be passed to the program: DATA: the name of the SAS input dataset; it is assumed that the input dataset was created with one line per time point per subject, and missing data points have been set to SAS missing data value.

OUTV: the name of the binary response variable
COVRS: the names of the explanatory variables, separated by spaces
NVARS: the total number of input variables (the ID identifier, the response variable and the explanatory variables)
NCOVS: the total number of explanatory variables; this is used only as a check

ID: the identifier field for the each subject; repeated data for the same subject will have the same ID.
YVAR: the index of the response variable, Y.
XVARS: indices for the explanatory variables, X.
VVAR: the 'working' correlation structure desired. Four correlation structures are allowed: (1) independent, (2) exchangeable, (3) one-dependent structure, (4) completely unspecified structure.
ITER: the total number of iterations desired, default is 20.
CRIT: the convergence criterion, default is 0.001.
PIN: the tail probability required for forward selection into model
POUT: the tail probability required for backward elimination from the model
SUBS: code for using mean imputation to replace missing elements (0 means exclude responses with missing covariates, 1 means replace missing covariate values with the mean for that field).

The program is invoked by passing the appropriate parameters in this manner after a data step has been defined to read in the input dataset DATA:

```
% STEPGEE (DATA,
          OUTV,
          COVRS,
          NVARS,
          NCOVS,
          ID,
          YVAR,
          XVARS,
          VVAR,
          ITER,
          CRIT,
          PIN,
          POUT,
          SUBS
          );
```

6. Example

We illustrate the above selection procedure with data on the functional decline in the activities of daily living (ADL) in a group of elderly subjects. A sample of 206 adults aged 75–100, who were part of a study on dysfunctional syndrome in the elderly were interviewed at five time periods: baseline information consisting of what the subject or a reliable surrogate could recall 2 weeks prior to admission to a local hospital for different kinds of ailments; at admission; at discharge; 30 days post-discharge and 90 days post-discharge. The outcome variable 'Decline' was defined as a decrease in the number of independent ADLs measured on five items (bathing, dressing, transfer, eating and toileting) from one time point to the next, resulting in four dependent binary responses for most subjects. Some had missing responses at certain times, so a total of 723 individual responses were obtained. The binary response variable, 'Decline', and the 17 covariables that the researchers deemed necessary are listed in Table 1. The following command was used to invoke STEPGEE. For example, the following data step part of SAS is also provided before the invocation of STEPGEE. A libname indicating the location of the input file is also included. A SAS dataset called elder.ssd was created from the sample data set.

```
LIBNAME run 'c:sas\sasuser\isaac';
DATA;
  SET run.elder;
RUN;
```

The STEPGEE macro is then invoked as:

```
%STEPGEE (run.elder,
          Decline,
          Sex Race Marital Age Adlbase
            Apache Depress Typefrom
          Destin Livespx Index LOS Health
            IADL Mobile QOL Neuroc,
          19,
          17,
          1,
          3,
          5 6 7 8 9 10 11 12 13 14 15 16 17
            18 19 20 21,
          20,
          0.001,
```

0.05,
0.10,
1);

The output results are listed in the Appendix. The stepwise selection procedure produced three significant variables at the PIN = 0.05 significant level. We assumed three correlation structures: the independent structure, the exchangeable or equal correlation structure and the completely unspecified structure. In this example, the same final model was selected. Only the results from the exchangeable correlation structure are therefore presented in the appendix. The goodness-of-fit statistics are included in the output. They show that the final model provided an adequate fit to the data for this type of model, since the contribution of variables not selected into the model could jointly be assumed to be zero ($P = 0.9983$). Under all correlation structures, the order of selection of all the variables was the same. The estimate of the common odds ratio is about 0.30 (from the log-odds of about −1.2), indicating a reduction in odds of decline with time. This reflects the fact that individuals who declined from one time point to the next were less likely to decline further after that.

7. Discussion

Many researchers have pointed out some of the inherent problems with any variable selection strategy (Harrel et al. [18], Miller [19]). Nevertheless, variable selection remains a valuable tool in data analysis. This paper represents the first attempt to apply stepwise regression procedures to a new and growing field. We have successfully implemented a stepwise procedure for correlated binary data. The performance of the selection strategy is currently being studied and will be reported elsewhere. Preliminary results suggest that under mild to moderate association, the choice of correlation structure does not adversely affect the selection criteria, that is, the same final model is usually selected, though the estimates of the regression coefficients might differ slightly. The appropriateness of the score tests used for selection was confirmed by low Type I errors which ranged from 3 to 9% (It was preset at 5%). We also compared the parameters estimates and robust standard errors obtained from STEPGEE with the results of two other items of software [5,7]. Using a sample dataset, referred to as visits.dat, provided by these authors in their examples, we dichotomiz-

Table 1

DECLINE	functional decline on binary scale 0 = no decline, 1 = decline
SEX	0 = male, 1 = female
RACE	0 = black, 1 = white
MARITAL	marital status 0 = married, 1 = not currently married
AGE	(range from 75 to 100)
ADLBASE	ADL at baseline (range from 0 to 5)
APACHE	apache score (range from 6 to 28)
DEPRESS	depression scale (range from 0 to 13)
TYPEFROM	type of residence from which subject was admitted 0 = home, 1 = other
DESTIN	residence to which subject will go after discharge 0 = home, 1 = other
LIVESPX	living situation before admission 0 = other, 1 = lived with spouse
INDEX	comorbid illness index (range from 0 to 10)
LOS	length of stay (days) range from 2 to 43 days
HEALTH	global health status 0 = good, 1 = poor
IADL	instrumental ADL of seven items (telephone, transportation, shopping, meals, housework, manage money, medications) (range from 0 to 7) a score of 7 indicates independence, i.e. do not need help on all seven items
MOBILE	mobility on four items (sit up in bed, write, walk across room, get from bedtochair) — range from 0 to 4
QOL	quality of life 0 = satisfied, 1 = dissatisfied,
NEUROC	neurocognitive assessment on a 21-mini mental state questionaire range from 0 to 21.

ed the response variable and modelled it using these items of software. In our case, we run STEPGEE with PIN = 0.95 and POUT = 0.96 to ensure that all variables were selected into the final model model. The same regression coefficients and standard errors were obtained by all three items of software.

From a data analytic point of view, the stepwise procedure implemented in the example, suggests that a patients assessment of global health, apprehension about where she/he will be discharged to after hospitalization, e.g. going home to family or nursing home, and loss of independence on instrumental ADLs seems to be most predictive of functional decline in activities of daily living. Although targeting hospitalized older patients according to risk for functional decline is often advocated, early predictors have not been previously identified and validated. Those who have attempted this have ignored the time dependence among the responses. Such a stepwise strategy may prove useful in identifying factors that ail these elderly patients at risk for functional decline, and thereby targeting them for adequate care.

8. Availability of program

A copy of the macro can be obtained by writing to Dr. Isaac F. Nuamah, University of Pennsylvania Cancer Center, 528 Blockley Hall, Philadelphia, PA 19101-6021, USA. Please include either a 3½ or a 5¼ inch diskette.

Acknowledgements

The authors wish to thank Seth Landefeld, M.D., of the Department of Internal Medicine, Case Western Reserve University, for making the data available to us. The data which came from a clinical trial on Dysfunctional Syndrome in Elderly Patients is funded by the John A. Hartford Foundation.

Appendix

```
     STEPWISE REGRESSION FOR CORRELATED BINARY DATA
     ==============================================

     Outcome Variable:   DECLINE
     Covariates:
SEX      RACE     MARITAL  AGE    ADLBASE  APACHE   DEPRESS
TYPEFROM  DESTIN   LIVESPX  INDEX    LOS    HEALTH   IADL
MOBILE   QOL   NEUROC

     Link :   LOGISTIC
     Correlation Structure:  EXCHANGEABLE
     Total Number of Records read:        723
     Total Number of Clusters:        206
     Maximum Cluster Size:          4
     Minimum Cluster Size:          1


B   0.0010229    (Step 0)


     CORRELATED BINARY REGRESSION MODEL
-------------------------------------------------------------
VARIABLE      ESTIMATE    ST.ERROR        T       P-VALUE
-------------------------------------------------------------
INTERCPT     -1.0522      0.0816    -12.8967      0.0000
SIGMA1       -2.0224      1.0183     -1.9860      0.0484
-------------------------------------------------------------
SIGMA 1  IS THE  LOG ODDS RATIO


     NAIVE SE     P

     0.0817   0.0000

COVARIATES NOT IN THE MODEL

COVARIATE    CHI-SQUARE    P-VALUE

SEX           0.0375       0.8465
RACE          0.0082       0.9278
MARITAL       1.8227       0.1770
AGE           2.1825       0.1396
ADLBASE       3.0599       0.0802
APACHE        0.4606       0.4973
DEPRESS       7.5638       0.0060
TYPEFROM      9.6377       0.0019
DESTIN       18.0198       0.0000
LIVESPX       1.6167       0.2036
INDEX         2.1815       0.1397
LOS           3.6430       0.0563
HEALTH       36.5624       0.0000
IADL         18.1898       0.0000
MOBILE        5.9854       0.0144
QOL           1.0826       0.2981
NEUROC        0.0017       0.9672

HEALTH    ENTERED THE MODEL

B   -0.000659 0.0023451   (Step 1)

     CORRELATED BINARY REGRESSION MODEL
-------------------------------------------------------------
VARIABLE      ESTIMATE    ST.ERROR        T       P-VALUE
-------------------------------------------------------------
INTERCPT     -1.9040      0.1870    -10.1800      0.0000
HEALTH        1.1979      0.2080      5.7588      0.0000
SIGMA1       -1.9085      0.9437     -2.0224      0.0444
-------------------------------------------------------------
SIGMA 1  IS THE  LOG ODDS RATIO


     NAIVE SE     P

     0.1932   0.0000
     0.2272   0.0000

COVARIATES NOT IN THE MODEL

COVARIATE    CHI-SQUARE    P-VALUE

SEX           0.0012       0.9719
RACE          0.0587       0.8086
MARITAL       0.4596       0.4978
AGE           3.2925       0.0696
ADLBASE       0.7412       0.3893
APACHE        0.1710       0.6792
```

```
DEPRESS        1.3701       0.2418
TYPEFROM       5.6124       0.0178
DESTIN        11.9371       0.0006
LIVESPX        0.4975       0.4806
INDEX          1.4189       0.2336
LOS            3.4785       0.0622
IADL          12.7487       0.0004
MOBILE         4.6266       0.0315
QOL            0.1790       0.6722
NEUROC         0.0180       0.8931

IADL      ENTERED THE MODEL



B    -0.00489 0.0036319  0.000832   (Step 2)

     CORRELATED BINARY REGRESSION MODEL
-------------------------------------------------------------
VARIABLE       ESTIMATE    ST.ERROR         T        P-VALUE
-------------------------------------------------------------
INTERCPT      -1.2583       0.2450     -5.1357       0.0000
HEALTH         1.0606       0.2055      5.1620       0.0000
IADL          -0.1605       0.0426     -3.7689       0.0002
SIGMA1        -1.2771       0.4214     -3.0306       0.0028
-------------------------------------------------------------
SIGMA 1  IS THE  LOG ODDS RATIO

     NAIVE SE      P

     0.2437   0.0000
     0.2256   0.0000
     0.0421   0.0001

COVARIATES NOT IN THE MODEL

COVARIATE     CHI-SQUARE     P-VALUE

SEX            0.0046       0.9462
RACE           0.4057       0.5241
MARITAL        0.6680       0.4138
AGE            1.7017       0.1921
ADLBASE        0.7084       0.4000
APACHE         0.0018       0.9665
DEPRESS        0.0000       0.9976
TYPEFROM       2.5397       0.1110
DESTIN         5.4606       0.0195
LIVESPX        0.7237       0.3949
INDEX          0.5504       0.4582
LOS            2.3745       0.1233
MOBILE         0.8836       0.3472
QOL            1.1721       0.2790
NEUROC         0.3130       0.5758

DESTIN     ENTERED THE MODEL

B   -0.005591 0.0040315 0.0007962 0.0010448 (Step 3)

     CORRELATED BINARY REGRESSION MODEL
-------------------------------------------------------------
VARIABLE       ESTIMATE    ST.ERROR         T        P-VALUE
-------------------------------------------------------------
INTERCPT      -1.5391       0.2883     -5.3385       0.0000
HEALTH         1.0122       0.2085      4.8551       0.0000
IADL          -0.1268       0.0477     -2.6590       0.0085
DESTIN         0.4151       0.1826      2.2728       0.0241
SIGMA1        -1.2196       0.3812     -3.1992       0.0016
-------------------------------------------------------------
SIGMA 1  IS THE  LOG ODDS RATIO

     NAIVE SE      P

     0.2747   0.0000
     0.2263   0.0000
     0.0445   0.0044
     0.1709   0.0152

COVARIATES NOT IN THE MODEL

COVARIATE     CHI-SQUARE     P-VALUE

SEX            0.0964       0.7562
RACE           1.0764       0.2995
MARITAL        0.0822       0.7743
AGE            0.7511       0.3861
ADLBASE        2.0036       0.1569
APACHE         0.0950       0.7579
DEPRESS        0.3530       0.5524
TYPEFROM       1.2223       0.2689
LIVESPX        0.1037       0.7474
INDEX          0.7609       0.3831
LOS            1.3372       0.2475
MOBILE         0.7003       0.4027
QOL            1.4249       0.2326
NEUROC         0.2758       0.5995

No Term passes the Enter(PIN) and Remove(POUT) Limits
 (  0.05 ,   0.10 )

Working Correlation Matrix

         1 -0.055312 -0.055312 -0.055312
-0.055312          1 -0.055312 -0.055312
-0.055312 -0.055312          1 -0.055312
-0.055312 -0.055312 -0.055312          1

Test of Model Adequacy:
GENERALIZED SCORE TEST  3.3365 P-value =  0.9983
```

References

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

[2] R.L. Prentice, Correlated binary regression with covariates specific to each binary observation, Biometrics 44 (1988) 1033–1048.

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

[4] SAS Institute INC. SAS/STAT User's Guide, Version 6, 4th edn., Vol. 2 (SAS Institute Inc, Cary, NC, 1989).

[5] M.R. Karim and S.L. Zeger, GEE: a SAS macro for longitudinal data analysis, Technical Report No. 674 (Department of Biostatistics, The Johns Hopkins University, 1988).

[6] SAS Institute INC. SAS/IML Software: Usage and Reference, Version 6 (SAS Institute Inc, Cary, NC, 1989).

[7] C.S. Davis, A computer program for regression analysis of repeated measures using generalized estimating equations, Comp. Methods Prog. Biomed. 40 (1993) 15–31.

[8] S.L. Zeger, K.Y. Liang and P. Albert, Models for longitudinal data: A generalized estimating equation approach, Biometrics 44 (1988) 1049–1060.

[9] K.Y. Liang, S.L. Zeger and B. Qaqish, Multivariate regression models for correlated categorical data (with discussion), J. R. Stat. Soc. B. 54 (1992) 3–40.

[10] S.R. Lipsitz, N.M. Laird and D.P. Harrington, Generalized estimating equations for correlated binary data: Use of the odds ratio as a measure of association, Biometrika 78 (1991) 153–160.

[11] K.V. Mardia, Some contributions to contigency-type bivariate distributions, Biometrika 54 (1967) 235–249.

[12] Y. Qu, G.W. Williams, G.J. Beck and S.V. Medendorp, Latent variables models for clustered dichotomous data with multiple subclasses, Biometrics 48 (1992) 1095–1102.

[13] V. Carey, S.L. Zeger and P. Diggle, Modelling multivariate binary data with alternating regression models, Biometrika 80 (1993) 517–526.

[14] C.R. Rao, Linear Statistical Inference and Its Applications, 2nd edn. (John Wiley & Sons, New York, 1973).

[15] D.D. Boos, On generalized score tests, Am. Stat. 46 (1992) 327–333.

[16] N. Breslow, Tests of hypothesis in overdispersed Poisson regression and other quasi-likelihood models, J. Am. Stat. Assoc. 85 (1990) 565–571.

[17] J.M. Krall, V.A. Uthoff and J.B. Harley, A step-up procedure for selecting variables associated with survival, Biometrics 31 (1975) 49–57.

[18] F.E. Harrel, K.L. Lee, R.M. Califf, D.B. Pryor and R.A. Rosati, Regression modelling strategies for improved prognostic prediction, Stat. Med. 3 (1984) 143–152.

[19] A.J. Miller, Subset selection in regression (Chapman and Hall, London, 1990).

[20] W.W. Hauck and R. Miike, A proposal for examining and reporting stepwise regressions, Stat. Med. 10 (1991) 711–715.