BIOMETRICS 51, 309–317
March 1995

SHORTER COMMUNICATIONS

EDITOR:
B. J. T. MORGAN

A Caveat Concerning Independence Estimating Equations With Multivariate Binary Data

Garrett M. Fitzmaurice

Department of Biostatistics, Harvard School of Public Health,
Boston, Massachusetts 02115, U.S.A.

SUMMARY

Clustered binary data occur commonly in both the biomedical and health sciences. In this paper, we consider logistic regression models for multivariate binary responses, where the association between the responses is largely regarded as a nuisance characteristic of the data. In particular, we consider the estimator based on independence estimating equations (IEE), which assumes that the responses are independent. This estimator has been shown to be nearly efficient when compared with maximum likelihood (ML) and generalized estimating equations (GEE) in a variety of settings. The purpose of this paper is to highlight a circumstance where assuming independence can lead to quite substantial losses of efficiency. In particular, when the covariate design includes within-cluster covariates, assuming independence can lead to a considerable loss of efficiency in estimating the regression parameters associated with those covariates.

1. Introduction

Clustered binary data occur commonly in both the biomedical and health sciences. The clustering may arise due to subsampling of the primary sampling unit, e.g., when observations are made on each member within a cluster or group. Alternatively, clustering can arise in longitudinal studies where repeated observations are made on the same unit across occasions. Whatever the nature of the clustering, observations within the same cluster are usually (positively) correlated. The focus of this paper is on regression models for multivariate binary responses, where the expectation of the response is related to a set of covariates by some known link function. When the responses are binary, a natural choice is to use a logit link function, although in principal any link functions can be used. We distinguish between studies where the association within the vector of responses is of scientific interest, from studies where the association parameters are considered to be a nuisance characteristic of the data. In the former, the parameters modelling the covariance or the *conditional* expectation given other responses are of primary interest, while in the latter interest is focussed primarily on the regression parameters for the *marginal* expectations. Regression or marginal models are the focus of this paper.

When the response is continuous and assumed to be approximately Gaussian, there is a general class of linear models that are suitable for analyses. However, when the response variable is binary, fewer methods are available. This is due in part to the lack of a discrete multivariate analogue of the multivariate Gaussian for the joint distribution of the responses. Thus, likelihood-based methods for multivariate binary outcomes have only been developed for certain special cases. Recently, there has been considerable interest in the generalized estimating equations (GEE) approach to analysing multivariate binary responses (Liang and Zeger, 1986; Zeger and Liang, 1986), which does not require the complete specification of the joint distribution of the responses. In particular, the estimator based on *independence* estimating equations (IEE), which assumes that the responses are independent, has been shown to be nearly efficient relative to maximum likelihood in a variety of settings. When the correlation between responses is not too high, Zeger (1988) suggests that this

Key words: Correlated binary responses; Generalized estimating equations; Marginal models.

estimator should be nearly efficient. In a more recent article that focuses on the bivariate case, McDonald (1993) concludes that the independence estimator "may be recommended for practical purposes whenever the association between paired observations is a nuisance." This estimator is also very appealing, since it can easily be implemented using existing statistical software packages. In contrast, Zhao, Prentice, and Self (1992) present asymptotic efficiency results that suggest that when the correlation between responses is high, assuming independence "can lead to important losses of efficiency." In this paper we demonstrate how the paradoxical findings of McDonald (1993) and Zhao, et al. (1992) can be explained by their choice of covariate design.

The purpose of this paper is to highlight a circumstance where assuming independence can lead to quite substantial losses of efficiency. In particular, when the covariate design includes a within-cluster covariate, that is, a covariate that varies *within* the cluster, assuming independence can lead to a considerable loss of efficiency in estimating the regression parameter associated with that covariate. Such covariates can arise in both designed experiments (e.g., higher-order cross-over designs for carryover effects (Baalam, 1968)) and observational studies. In longitudinal studies, within-cluster covariates are often called "time-varying" covariates, since they can change from one occasion of measurement to another. In Section 2 we introduce some notation and discuss the marginal distribution of the vector of binary responses. Estimators of the marginal regression parameters, based on independence and generalized estimating equations are described in Section 3. A full likelihood-based estimator is also described. In Section 4, we compare the asymptotic efficiency of the estimator based on independence estimating equations relative to the optimal maximum likelihood estimator. Finally, a tractable expression for the limiting asymptotic relative efficiency of the "independence" estimator relative to the GEE estimator is derived.

2. Notation

Consider N multivariate binary observations $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_N$, where $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{im_i})'$, and $i = 1, 2, \dots, N$ indexes clusters. Suppose that each cluster has a $Q \times 1$ covariate vector $\mathbf{x}_{ij}$ associated with Y_{ij}, where $j = 1, 2, \dots, m_i$ indexes units within a cluster. Letting $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{im_i})'$ represent the $m_i \times Q$ matrix of covariates for the ith cluster, the marginal distribution of Y_{ij} is Bernoulli,

$$f(y_{ij}|\mathbf{X}_i) = \exp[y_{ij}\theta_{ij} - \log\{1 + \exp(\theta_{ij})\}],$$

where we assume $\theta_{ij} = \log\{\mu_{ij}/(1 - \mu_{ij})\} = \mathbf{x}_{ij}'\boldsymbol{\beta}$, and $\mu_{ij}(\boldsymbol{\beta}) = E(Y_{ij}) = \text{pr}(Y_{ij} = 1|\mathbf{x}_{ij}, \boldsymbol{\beta})$ is the probability of success on Y_{ij}; and $\boldsymbol{\beta}$ is a $Q \times 1$ vector of regression parameters. With binary responses, the logit link function is a natural choice although, in principle, any link function could be chosen. The $\mu_{ij}(\boldsymbol{\beta})$ can be grouped together to form a vector $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ containing the marginal probabilities of success, $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = E(\mathbf{Y}_i) = (\mu_{i1}, \dots, \mu_{im_i})'$. In the preceding, the only assumption made concerns the marginal distribution of Y_{ij}. In the next section, we consider estimators of $\boldsymbol{\beta}$ based on independence and generalized estimating equations. A likelihood-based estimator of $\boldsymbol{\beta}$ is also described.

3. Estimating $\boldsymbol{\beta}$

3.1 *Independence Estimating Equations*

If the responses are naively assumed to be independent, then their joint distribution is

$$f(\mathbf{y}_i|\mathbf{X}_i) = \exp\left[\sum_{j=1}^{mi} y_{ij}\theta_{ij} - \sum_{j=1}^{mi} \log\{1 + \exp(\theta_{ij})\}\right].$$

The maximum likelihood estimate, $\hat{\boldsymbol{\beta}}_I$, is the solution to the independence estimating equations,

$$\sum_{i=1}^{N} \frac{\partial l_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \mathbf{X}_i'\boldsymbol{\Delta}_i\text{cov}^{-1}(\mathbf{Y}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i) = \sum_{i=1}^{N} \mathbf{X}_i'(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0, \qquad (1)$$

where $\boldsymbol{\Delta}_i = \text{diag}[\text{var}(Y_{i1}), \dots, \text{var}(Y_{im_i})]$.

In spite of the fact that the responses on units within the same cluster are usually (positively) correlated, ordinary logistic regression maximum likelihood estimation (which assumes the within-cluster responses are independent) yields estimates which are consistent and asymptotically normal (Liang and Zeger, 1986). However, the joint likelihood under independence ignores the possible intra-cluster correlation among the binary responses. Consequently, the inverse of the estimated information matrix can give inconsistent estimates of the asymptotic variance of the estimated regression parameters. To circumvent this problem, Liang and Zeger propose using a robust estimate of the variance, which is consistent regardless of the true correlation between the responses. This robust variance was also proposed by Huber (1967), and more recently by White (1982), and Royall (1986), and is described in the next section. Thus, one very simple approach to analysing multivariate binary responses is to use ordinary logistic regression, followed by a robust variance correction.

3.2 *Generalized Estimating Equations*

With a binary response vector, the generalized estimating equations developed by Liang and Zeger (1986) and Prentice (1988), simply generalize the independence estimating equations given in (1) by introducing a "working" or approximate correlation matrix, $\mathcal{R}_i(\alpha)$. This leads to estimating equations of the form,

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbf{X}_i'\boldsymbol{\Delta}_i\mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0, \tag{2}$$

where $\mathbf{V}_i = \boldsymbol{\Delta}_i^{1/2}\mathcal{R}_i(\boldsymbol{\alpha})\boldsymbol{\Delta}_i^{1/2}$. These generalized estimating equations yield consistent estimators of the regression parameters, under only the correct specification of the form of the mean function, $\boldsymbol{\mu}_i$. Since $\boldsymbol{\alpha}$ is unknown, it is usually estimated by defining a $m_i(m_i - 1)/2$ vector of empirical correlations, r_i, with elements,

$$r_{ijk} = \frac{(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})}{\sqrt{\mu_{ij}(1 - \mu_{ij})\mu_{ik}(1 - \mu_{ik})}},$$

and then using a second set of moment estimating equations similar to (2). In many cases, this leads to simple noniterative methods for estimating $\boldsymbol{\alpha}$. Furthermore, the choice of estimator for $\boldsymbol{\alpha}$ has no effect on the asymptotic efficiency for estimating $\boldsymbol{\beta}$ (Newey, 1990). Thus, in general, the estimates $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ can be obtained by iterating between a modified Fisher scoring algorithm for $\boldsymbol{\beta}$ and moment estimation of $\boldsymbol{\alpha}$.

Finally, if the association between responses has been correctly specified, so that $\mathbf{V}_i = \text{cov}(\mathbf{Y}_i)$, then $\hat{\boldsymbol{\beta}}$ normalized has asymptotic covariance matrix given by,

$$\lim_{N\to\infty} N[H_1^{-1}(\boldsymbol{\beta})], \tag{3}$$

where

$$H_1(\boldsymbol{\beta}) = \sum_{i=1}^{N} (\mathbf{X}_i'\boldsymbol{\Delta}_i\mathbf{V}_i^{-1}\boldsymbol{\Delta}_i\mathbf{X}_i).$$

A consistent estimate of the asymptotic covariance of $\hat{\boldsymbol{\beta}}$ is given by $H_1^{-1}(\hat{\beta})$. However, if the "working" correlation, $\mathcal{R}_i(\boldsymbol{\alpha})$, is misspecified, then $\hat{\boldsymbol{\beta}}$ normalized has asymptotic covariance matrix given by,

$$\lim_{N\to\infty} N[H_1^{-1}(\boldsymbol{\beta})H_2(\boldsymbol{\beta})H_1^{-1}(\boldsymbol{\beta})], \tag{4}$$

where

$$H_2(\boldsymbol{\beta}) = \sum_{i=1}^{N} [\mathbf{X}_i'\boldsymbol{\Delta}_i\mathbf{V}_i^{-1}\text{cov}(\mathbf{Y}_i)\mathbf{V}_i^{-1}\boldsymbol{\Delta}_i\mathbf{X}_i].$$

A consistent estimate of the asymptotic covariance of $\hat{\boldsymbol{\beta}}$ is given by $H_1^{-1}(\hat{\boldsymbol{\beta}})H_2(\hat{\boldsymbol{\beta}})H_1^{-1}(\hat{\boldsymbol{\beta}})$, where

$$H_2(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{N} [\mathbf{X}_i'\hat{\boldsymbol{\Delta}}_i\hat{\mathbf{V}}_i^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)'\hat{\mathbf{V}}_i^{-1}\hat{\boldsymbol{\Delta}}_i\mathbf{X}_i].$$

This estimator is robust in the sense of being consistent even if the "working" covariance, $\mathbf{V}_i$, is not equal to $\text{cov}(\mathbf{Y}_i)$ (Liang and Zeger, 1986). Finally, we note that Zhao and Prentice (1990) and Liang, Zeger, and Qaqish (1992) have described extensions of the GEE methodology to allow for joint estimation of the mean and association parameters. The main advantage of the latter approaches is that they lead to more efficient estimates of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, provided the model for both the mean and the association is correctly specified. However, a serious drawback is that $\hat{\boldsymbol{\beta}}$ may fail to be consistent when the model for the association is misspecified.

3.3 *Maximum Likelihood Estimation*

A likelihood-based approach requires the complete representation of the joint probabilities of the vector of binary responses for each cluster. This joint distribution is multinomial with a vector of $2^{m_i} - 1$ non-redundant probabilities, $\boldsymbol{\pi}_i = E(\mathbf{P}_i)$, where

$$\mathbf{P}_i = (Y_{i1}Y_{i2}\cdots Y_{i(m_i-1)}Y_{im_i}, (1-Y_{i1})Y_{i2}\cdots Y_{i(m_i-1)}Y_{m_i}, \ldots,$$

$$(1-Y_{i1})(1-Y_{i2})\cdots(1-Y_{i(m_i-1)})Y_{im_i}).$$

The fully parameterized distribution has $2^{m_i} - 1$ parameters. In this section we outline one parametric description of the joint distribution, that was first suggested by Bahadur (1961), and later by Cox (1972).

Bahadur (1961) describes the joint distribution in terms of the marginal means, $\boldsymbol{\mu}_i$, and the marginal correlations, $\boldsymbol{\rho}_i = (\rho_{i12}, \rho_{i13}, \ldots, \rho_{i12\cdots m_i})$. Bahadur's representation of the joint distribution for $\mathbf{Y}_i$ can be written as,

$$f(\mathbf{y}_i|\boldsymbol{\mu}_i, \boldsymbol{\rho}_i) = \prod_{j=1}^{mi} \mu_{ij}^{y_{ij}}(1-\mu_{ij})^{1-y_{ij}}$$

$$\cdot\left(1 + \sum_{j<k} \rho_{ijk}e_{ij}e_{ik} + \sum_{j<k<l} \rho_{ijkl}e_{ij}e_{ik}e_{il} + \cdots + \rho_{i12\cdots m_i}e_{i1}e_{i2}\cdots e_{im_i}\right),$$

where $e_{ij} = \dfrac{(Y_{ij} - \mu_{ij})}{\sqrt{\mu_{ij}(1-\mu_{ij})}}$; and $\rho_{ijk} = E(e_{ij}e_{ik}), \ldots, \rho_{i12\cdots m_i} = E(e_{i1}e_{i2}\cdots e_{im_i})$. Thus, in terms of the $(2^{m_i} - m_i - 1)$ *marginal* correlations, $\boldsymbol{\rho}_i = (\rho_{i12}, \rho_{i13}, \ldots, \rho_{i12\cdots m_i})$, the joint distribution of the responses can be evaluated in closed form. A feature of Bahadur's representation, however, is that the marginal correlations must satisfy certain linear inequalities determined by the marginal probabilities. That is, the marginal correlations are constrained by the marginal probabilities.

The marginal probabilities, $\mu_{ij}(\boldsymbol{\beta})$, can be modelled in the usual way as a function of x_{ij} and $\boldsymbol{\beta}$. Then, one model for the pairwise and higher-way correlations, $\boldsymbol{\rho}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = (\rho_{i12}, \rho_{i13}, \ldots, \rho_{i12\cdots m_i})$, is to assume that they do not depend on cluster-level covariates, i.e., $\boldsymbol{\rho}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = (\rho_{12}, \rho_{13}, \ldots, \rho_{12\cdots m_i})$. Maximum likelihood estimates of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ can then be obtained by setting the first derivative of the log-likelihood with respect to $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ to zero and solving for $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$. The likelihood equations for $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ are,

$$S(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} S_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \mathbf{C}_i'\mathbf{W}_i^{-1}(\mathbf{P}_i - \boldsymbol{\pi}_i) = 0, \tag{5}$$

where $\mathbf{C}_i = \dfrac{\partial\boldsymbol{\pi}_i}{\partial(\boldsymbol{\beta},\boldsymbol{\alpha})}$, and $\mathbf{W}_i = \{\mathrm{diag}(\boldsymbol{\pi}_i)\} - \boldsymbol{\pi}_i\boldsymbol{\pi}_i'$. A more detailed account of the derivations can be found in Lipsitz, Laird, and Harrington (1990), who derive a set of likelihood equations similar to (5) for the bivariate case. The maximum likelihood estimates of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ can be obtained using a Newton–Raphson algorithm. Finally, if the model has been correctly specified, $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ normalized has asymptotic covariance matrix given by,

$$\lim_{N\to\infty} N\left(\sum_{i=1}^{N} \mathbf{C}_i'\mathbf{W}_i^{-1}\mathbf{C}_i\right)^{-1}. \tag{6}$$

4. Asymptotic Relative Efficiency

In this section, we examine the asymptotic relative efficiency of GEE estimators of the regression parameters, $\boldsymbol{\beta}$, when compared to the ML estimator. We assume that the mean structure has been correctly specified, but allow that the association between responses may be incorrectly specified, e.g. by incorrectly assuming *independence*. We consider clustered binary data arising from a simple longitudinal design with a *trivariate* response and a two-group design.

One way to calculate the asymptotic relative efficiency is to specify the true joint distribution of the binary responses and the covariates. That is, the probability of ($\mathbf{Y}_i$, $\mathbf{X}_i$) can be written as,

$$\Pr(\mathbf{Y}_i, \mathbf{X}_i) = \Pr(\mathbf{Y}_i|\mathbf{X}_i)\Pr(\mathbf{X}_i)$$

and is then fully determined by specifying:

1. The probability distribution of the covariates, $\mathbf{X}_i$.
2. The model for the probability of $\mathbf{Y}_i$ given $\mathbf{X}_i$, and ($\boldsymbol{\beta}$, $\boldsymbol{\alpha}$).

For specifying 1 and 2 above, we assume the following model for the marginal expectation of $\mathbf{Y}_i$ given $\mathbf{X}_i$,

$$\text{logit}(\mu_{it}) = \beta_0 + \beta_1 x_{it} + \beta_2(t-2); \qquad t = 1, 2, 3;$$

where x_{it} is a dichotomous covariate indicating group membership for the ith observation at the tth occasion. We consider two different design configurations, one where group is a cluster-level covariate (i.e., $x_{i1} = x_{i2} = x_{i3}$) and the other where group is a within-cluster or time-varying covariate. For both cases, each of the possible covariate configurations is assumed to have equal probability of occurrence. Note, we are assuming that x_{it} is an *external* time-varying covariate in the sense described by Kalbfleisch and Prentice (1980); otherwise, the maximum likelihood and GEE estimators of the parameters of the marginal model may be inconsistent (A discussion of this point can be found in Fitzmaurice, Laird, and Rotnitzky, 1993; and Pepe and Anderson, 1994). Finally, the association between the binary responses is parameterized in terms of the marginal correlations, $\rho_i = (\rho_{i12}, \rho_{i13}, \rho_{i23}, \rho_{i123})$.

The parameters of the true model are

$$\beta_0 = .25;\ \beta_1 = -.25;\ \beta_2 = -.25;$$

$$\rho_i = (\rho_{i12}, \rho_{i13}, \rho_{i23}, \rho_{i123}) = (\rho, \rho^2, \rho, 0); \qquad \text{where } \rho \in (0, .60).$$

That is, we assume the correlation between Y_{ij} and Y_{ik} is of the form $\rho^{|j-k|}, j \neq k$. This corresponds to the *first-order autoregressive* correlation pattern so commonly assumed for continuous time series data. We examine the asymptotic relative efficiency (ARE) of the estimator based on either independence or generalized estimating equations (assuming common pairwise correlations) relative to the maximum likelihood estimator, for both the group and time effects. The asymptotic relative efficiency (ARE) for any element of $\boldsymbol{\beta}$ is given by the ratio of the corresponding diagonal elements of (6) and (4). Results are not reported for the intercept term, since it is usually regarded as a nuisance parameter in this setting.

First we consider the case where group is a cluster-level covariate. The ARE for the time effect is very close to 1.0, regardless of the strength of the correlation between responses. However, for the group effect, there is a very discernible loss of efficiency for both the "independence" and "pairwise" GEE estimators. In Figure 1(a), the ARE for the group effect is plotted against the correlation parameter, ρ. The efficiency of both estimators declines with increasing correlation, and the decline is most notable when the correlation is greater than .4. The decline in efficiency can be explained by the fact that both the "independence" and "pairwise" GEE estimators fail to exploit all the information about the mean or regression parameters in the second and third moment parameters. Recall that with binary responses the mean parameters are constrained by the higher order moment parameters, and vice versa (see Prentice, 1988). Note, however, that in terms of their asymptotic efficiency, the "independence" and "pairwise" GEE estimators are almost indistinguishable. That is, the ARE of the "independence" estimator relative to the "pairwise" GEE estimator is almost unity. Next we consider the case where group is a within-cluster or time-varying covariate. Once again, the ARE for the time effect is very close to 1.0, regardless of the strength of the correlation between responses. However, for the group effect, there is a substantial loss of efficiency. Note, in Figure 1(b), that the efficiency of the "pairwise" GEE estimator can drop as low as 60%, and that there is a more notable loss of efficiency for the "independence" estimator. The efficiency of the "independence" estimator decreases quite rapidly with increasing correlation between the responses.

Comparing the "independence" estimator relative to the "pairwise" GEE estimator, the results in Figure 1(a) and 1(b) demonstrate that their ARE depends on the covariate design. That is, when the covariate design only includes cluster-level covariates the ARE is very close to 1.0. However, when the covariate design includes a within-cluster covariate, assuming independence can lead to a considerable loss of efficiency in estimating the regression parameter corresponding to that cova-

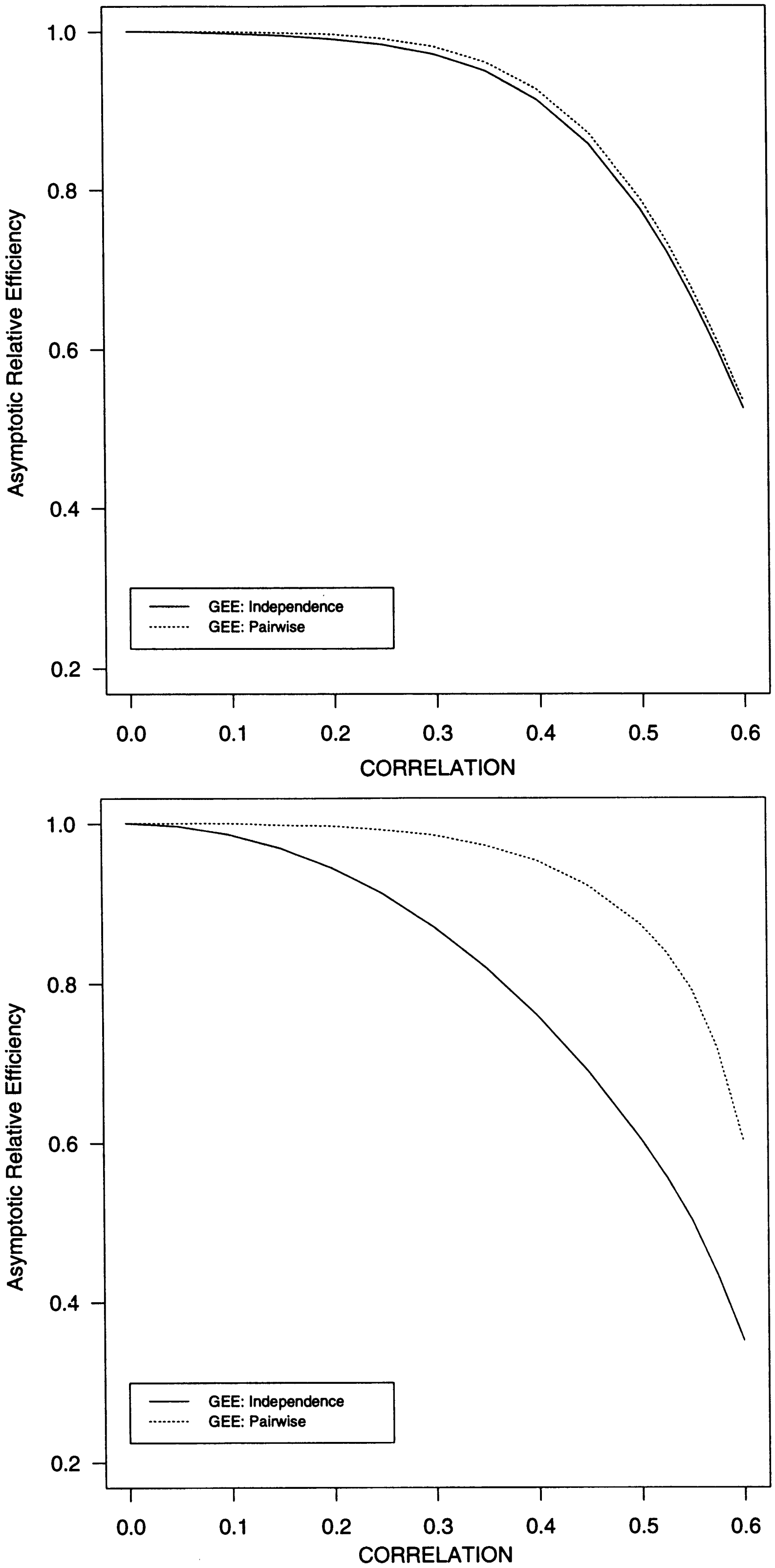

Figure 1. (a) Asymptotic efficiency of the GEE estimators, relative to the maximum likelihood estimator, when the true underlying joint distribution has a Bahadur representation (cluster-level covariate). (b) Asymptotic efficiency of the GEE estimators, relative to the maximum likelihood estimator, when the true underlying joint distribution has a Bahadur representation (within-cluster covariate).

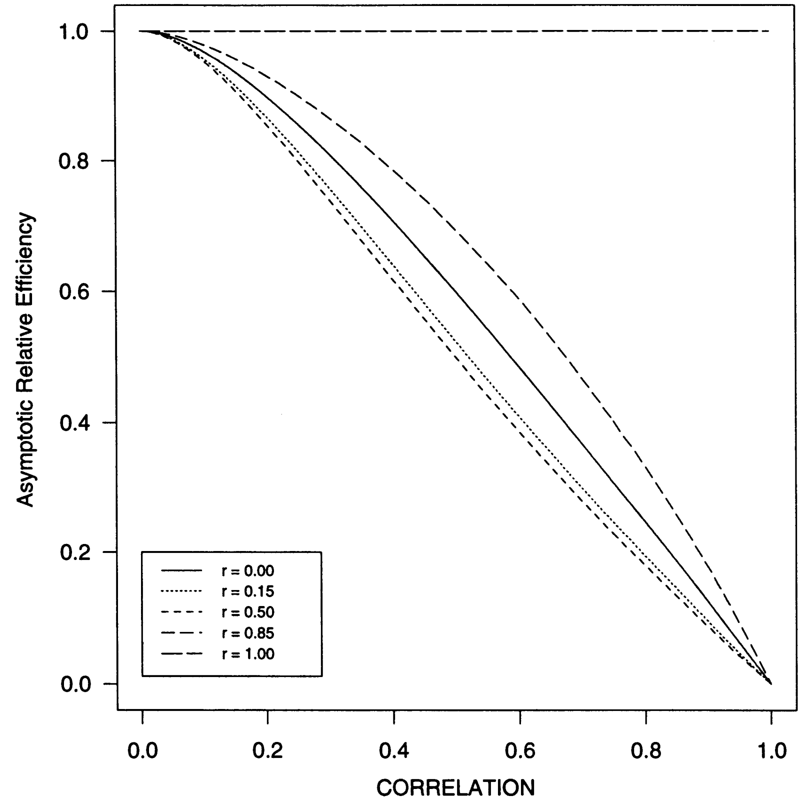

Figure 2. Asymptotic efficiency of the Exchangeable GEE estimator, relative to the Independence estimator, for selected values of the intra-cluster correlation for the covariate (and when $T = 5$).

riate. Finally, note that these two different covariate designs represent two extremes, one where the intra-cluster correlation between thex_{it} values is 1.0 (cluster-level covariate) and the other where the intra-cluster correlation is zero (within-cluster covariate, with each covariate configurations having an equal probability of occurrence). In the next paragraph, we explore in more detail how the ARE depends on *both* the intra-cluster correlation for the covariate (i.e., $\text{corr}(x_{it}, x_{it'})$, $t \neq t'$) and the correlation between the responses.

Following the approaches of Zhao, et al. (1992) and Lee, Scott, and Soo (1993), we consider the simple case of a scalar parameter, i.e., where there is only a single covariate (and no intercept). We assume the following model for the marginal expectation of $\mathbf{Y}_i$ given $\mathbf{X}_i$,

$$\text{logit}(\mu_{it}) = \beta x_{it}; \qquad t = 1, 2, \ldots, T.$$

Furthermore, we assume that $\rho_{ist} = \rho$, $s \neq t$, and arbitrarily fix the higher-way correlations to zero. Next, we assume that $\mathbf{X}_i$ has a multivariate normal distribution, with zero mean vector, and positive definite covariance matrix $\mathbf{\Sigma}_i = \{(1 - r)\mathbf{I} + r\mathbf{J}\}$, where $\mathbf{J}$ is a $T \times T$ matrix of ones and $\mathbf{I}$ is a $T \times T$ identity matrix. Then the asymptotic relative efficiency of the "independence" estimator relative to the "exchangeable" GEE estimator is

$$\text{ARE} = \lim_{N\to\infty} \left\{ \frac{\left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{\Delta}_i\mathbf{V}_i^{-1}\text{cov}(\mathbf{Y}_i)\mathbf{V}_i^{-1}\mathbf{\Delta}_i\mathbf{X}_i\right)\left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{\Delta}_i\mathbf{X}_i\right)^2}{\left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{\Delta}_i\mathbf{V}_i^{-1}\mathbf{\Delta}_i\mathbf{X}_i\right)^2\left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{\Delta}_i\text{cov}(\mathbf{Y}_i)\mathbf{\Delta}_i\mathbf{X}_i\right)} \right\}$$

$$= \lim_{N\to\infty} \left\{ \frac{\left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{\Delta}_i\mathbf{X}_i\right)^2}{\left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{\Delta}_i\text{cov}^{-1}(\mathbf{Y}_i)\mathbf{\Delta}_i\mathbf{X}_i\right)\left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{\Delta}_i\text{cov}(\mathbf{Y}_i)\mathbf{\Delta}_i\mathbf{X}_i\right)} \right\};$$

since $\mathbf{V}_i = \text{cov}(\mathbf{Y}_i)$ when the association between responses has been correctly specified.

In order to derive a tractable expression for the limiting ARE, we consider the special case where $\boldsymbol{\beta} = 0$, and thus $\boldsymbol{\Delta}_i = \frac{1}{2}\mathbf{I}$. Then, by the strong laws of large numbers, as $N \to \infty$,

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i'\boldsymbol{\Delta}_i\mathbf{X}_i \xrightarrow{wp1} \frac{T}{4};$$

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i'\boldsymbol{\Delta}_i\text{cov}(\mathbf{Y}_i)\boldsymbol{\Delta}_i\mathbf{X}_i \xrightarrow{wp1} \frac{T}{4}(1 + (T-1)\rho r);$$

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i'\boldsymbol{\Delta}_i\text{cov}^{-1}(\mathbf{Y}_i)\boldsymbol{\Delta}_i\mathbf{X}_i \xrightarrow{wp1} \frac{T}{4}\frac{(1 + (T-2)\rho - (T-1)\rho r)}{(1-\rho)(1+(T-1)\rho)}.$$

Thus, as $N \to \infty$,

$$\text{ARE} \xrightarrow{wp1} \frac{(1-\rho)(1+(T-1)\rho)}{(1+(T-1)\rho r)(1+(T-2)\rho-(T-1)\rho r)}.$$

This limit is a multivariate generalization of the one derived by Lee, et al. (1993) for the bivariate case and can also be obtained from the covariance expressions derived by Neuhaus (1993). Note that when either $\rho = 0$ or $r = 1.0$, the ARE is unity. That is, the ARE is 1.0 when either the responses are independent or $\mathbf{X}_i$ is a cluster-level covariate. However, if $\mathbf{X}_i$ is a within-cluster covariate ($r <$ 1.0), then the ARE declines with increasing correlation (ρ) between the responses. Note that these results explain the paradoxical findings of McDonald (1993) and Zhao, et al. (1992). McDonald (1993) considered a covariate design with only cluster-level covariates ($r = 1.0$), and thus found that assuming independence results in little loss of efficiency. On the other hand, Zhao, et al. (1992) considered a covariate design with a within-cluster covariate and $r = 0.0$, and thus found that assuming independence results in substantial losses of efficiency. Finally, we note that for fixed ρ, the ARE does not decrease monotonically with decreasing r. This is demonstrated in Figure 2, where the ARE is plotted against ρ for selected values of the intra-cluster correlation for the covariate (r) (and when $T = 5$).

5. Conclusion

In this paper we consider estimators of the logistic regression parameters in models for multivariate binary responses. In particular, we consider the estimator based on independence estimating equations, which assumes that the responses are independent. This estimator has been shown to be nearly efficient when compared with maximum likelihood and generalized estimating equations in a variety of settings. Thus, one very simple approach to analysing multivariate binary responses is to use ordinary logistic regression, followed by a robust variance correction. We show, however, that when the responses are strongly correlated and the covariate design includes a within-cluster covariate, assuming independence can lead to a considerable loss of efficiency in estimating the regression parameter associated with that covariate. This result demonstrates that the degree of efficiency depends on *both* the strength of the correlation between the responses *and* the covariate design. Finally, although these results are asymptotic, Lee, et al. (1993) and Lipsitz, et al. (1994) report similar findings from a simulation study with small to moderate sample sizes. In conclusion, we recommend that some attempt should generally be made to model the association between responses, even when the association is regarded as a nuisance characteristic of the data and its precise nature is unknown.

Acknowledgements

The author thanks the Editor and two referees for their helpful comments and suggestions on an earlier version of the paper. This research was supported by grants GM29745 and MH17119 from the National Institutes of Health.

Résumé

Des données binaires agrégées en classes sont fréquentes à la fois dans les sciences biomédicales et de la santé. Dans cet article, nous considérons les modèles de régression logistique dans le cas de réponses binaires multidimensionnelles, oú l'association entre les réponses est considérée pour unc grande part comme une caractéristique de nuisance des connées. Nous considérons en particulier l'estimateur basé sur les équations d'estimation indépendantes (IEE), qui suppose que les réponses sont indépendantes.

Cet estimateur est jugé comme étant á peu près efficace par rapport à la méthode du maximum de vraisemblance (ML) et aux équations d'estimation généralisées (GEE) dans une variété de contextes. Le but de cet article est de présenter une situation oú l'hypothése d'indépendance peut conduire à une perte substantielle d'efficacité. En particulier lorsque les covariables du dispositif comprennent des covariables intra-classes, l'indépendance peut entraîner une perte considérable d'efficacité dans l'estimation des coefficients de régression associés à ces covariables.

References

Baalam, L. N. (1968). A two-period design with t^2 experimental units. *Biometrics* **24,** 61–73.

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

Cox, D. R. (1972). The analysis of multivariate binary data. *Applied Statistics* **21,** 113–120.

Fitzmaurice, G. M., Laird, N. M., and Rotnitzky, A. G. (1993). Regression models for discrete longitudinal responses (with discussion). *Statistical Science* **8(3),** 248–309.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1, 221–233. Berkeley: University of California Press.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley.

Lee, A. J., Scott, A. J., and Soo, S. C. (1993). Comparing Liang-Zeger estimates with maximum likelihood in bivariate logistic regression. *Journal of Statistical Computation and Simulation* **44,** 133–148.

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

Liang, K. Y., Zeger, S. L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal Statistical Society, Series B* **54,** 3–40.

Lipsitz, S. R., Fitzmaurice, G. M., Orav, E. J., and Laird, N. M. (1994). Performance of generalized estimating equations in practical situations. *Biometrics* **50,** 270–278.

Lipsitz, S. R., Laird, N. M., and Harrington, D. P. (1990). Maximum likelihood regression methods for paired binary data. *Statistics in Medicine* **9,** 1517–1525.

McDonald, B. W. (1993). Estimating logistic regression parameters for bivariate binary data. *Journal of the Royal Statistical Society, Series B* **55,** 391–397.

Neuhaus, J. M. (1993). Estimation efficiency and test of covariate effects with clustered binary data. *Biometrics* **49,** 989–996.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5,** 99–135.

Pepe, M. S. and Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics* **23,** 939–951.

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

Royall, R. M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review* **54,** 221–226.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50,** 1–26.

Zeger, S. L. (1988). The analysis of discrete longitudinal data: Commentary. *Statistics in Medicine* **7,** 161–168.

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

Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77,** 642–648.

Zhao, L. P., Prentice, R. L., and Self, S. G. (1992). Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society, Series B* **54,** 805–811.

Received October 1993; *final revision August* 1994; *accepted September* 1994.