Biometrika (1986), **73**, 1, *pp.* 13–22
Printed in Great Britain

Longitudinal data analysis using generalized linear models

BY KUNG-YEE LIANG AND SCOTT L. ZEGER
Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.

SUMMARY

This paper proposes an extension of generalized linear models to the analysis of longitudinal data. We introduce a class of estimating equations that give consistent estimates of the regression parameters and of their variance under mild assumptions about the time dependence. The estimating equations are derived without specifying the joint distribution of a subject's observations yet they reduce to the score equations for multivariate Gaussian outcomes. Asymptotic theory is presented for the general class of estimators. Specific cases in which we assume independence, m-dependence and exchangeable correlation structures from each subject are discussed. Efficiency of the proposed estimators in two simple situations is considered. The approach is closely related to quasi-likelihood.

Some key words: Estimating equation; Generalized linear model; Longitudinal data; Quasi-likelihood; Repeated measures.

1. INTRODUCTION

Longitudinal data sets, comprised of an outcome variable, y_{it}, and a $p \times 1$ vector of covariates, x_{it}, observed at times $t = 1, \ldots, n_i$ for subjects $i = 1, \ldots, K$ arise often in applied sciences. Typically, the scientific interest is either in the pattern of change over time, e.g. growth, of the outcome measures or more simply in the dependence of the outcome on the covariates. In the latter case, the time dependence among repeated measurements for a subject is a nuisance. For example, the severity of respiratory disease along with the nutritional status, age, sex and family income of children might be observed once every three months for an 18 month period. The dependence of the outcome variable, severity of disease, on the covariates is of interest.

With a single observation for each subject ($n_i = 1$), a generalized linear model (McCullagh & Nelder, 1983) can be applied to obtain such a description for a variety of continuous or discrete outcome variables. With repeated observations, however, the correlation among values for a given subject must be taken into account. This paper presents an extension of generalized linear models to the analysis of longitudinal data when regression is the primary focus.

When the outcome variable is approximately Gaussian, statistical methods for longitudinal data are well developed, e.g. Laird & Ware (1982) and Ware (1985). For non-Gaussian outcomes, however, less development has taken place. For binary data, repeated measures models in which observations for a subject are assumed to have exchangeable correlations have been proposed by Ochi & Prentice (1984) using a probit link, by Stiratelli, Laird & Ware (1984) using a logit link and by Koch et al. (1977) using log linear models. Only the model proposed by Stiratelli, Laird & Ware allows for time-

chain model for binary longitudinal data which, also, however, requires time independent covariates. One difficulty with the analysis of non-Gaussian longitudinal data is the lack of a rich class of models such as the multivariate Gaussian for the joint distribution of y_{it} $(t = 1,\ldots,n_i)$. Hence likelihood methods have not been available except in the few cases mentioned above.

The approach in this paper is to use a working generalized linear model for the marginal distribution of y_{it}. We do not specify a form for the joint distribution of the repeated measurements. Instead, we introduce estimating equations that give consistent estimates of the regression parameters and of their variances under weak assumptions about the joint distribution. We model the marginal rather than the conditional distribution given previous observations although the conditional approach may be more appropriate for some problems. The methods we propose reduce to maximum likelihood when the y_{it} are multivariate Gaussian.

The estimating equations introduced here are similar to those described by Jorgensen (1983) and by Morton (1981). However our problem differs from the one considered by Jorgensen in that the correlation parameters do not appear in the estimating equations in an additive way; it is different than the problem considered by Morton in that pivots cannot be used to remove the nuisance correlation parameters.

To establish notation, we let $Y_i = (y_{i1},\ldots,y_{in_i})^{\mathrm{T}}$ be the $n_i \times 1$ vector of outcome values and $X_i = (x_{i1},\ldots,x_{in_i})^{\mathrm{T}}$ be the $n_i \times p$ matrix of covariate values for the ith subject $(i = 1,\ldots,K)$. We assume that the marginal density of y_{it} is

$$f(y_{it}) = \exp\left[\{y_{it}\theta_{it} - a(\theta_{it}) + b(y_{it})\}\,\phi\right], \tag{1}$$

where $\theta_{it} = h(\eta_{it})$, $\eta_{it} = x_{it}\beta$. By this formulation, the first two moments of y_{it} are given by

$$E(y_{it}) = a'(\theta_{it}), \quad \mathrm{var}\,(y_{it}) = a''(\theta_{it})/\phi. \tag{2}$$

When convenient to simplify notation, we let $n_i = n$ without loss of generality.

Section 2 presents the 'independence' estimating equation which arises by adopting the working assumption that repeated observations for a subject are independent. It leads to consistent estimates of β and of its variance given only that the regression model for $E(y)$ is correctly specified. Section 3 introduces and presents asymptotic theory for the 'generalized' estimating equation in which we borrow strength across subjects to estimate a 'working' correlation matrix and hence explicitly account for the time dependence to achieve greater asymptotic efficiency. In § 4, examples of specific models to be used in the analysis of longitudinal data are given. Section 5 considers questions of efficiency. The final section discusses several issues concerning the use of these estimating procedures.

2. INDEPENDENCE ESTIMATING EQUATIONS

In this section, we present an estimator, $\hat{\beta}_I$, of β which arises under the working assumption that repeated observations from a subject are independent of one another. Under the independence working assumption, the score equations from a likelihood analysis have the form

$$U_I(\beta) = \sum_{i=1}^{K} X_i^{\mathrm{T}} \Delta_i S_i = 0, \tag{3}$$

where $\Delta_i = \mathrm{diag}\,(d\theta_{it}/d\eta_{it})$ is an $n \times n$ matrix and $S_i = Y_i - a_i'(\theta)$ is of order $n \times 1$ for the ith subject. The estimator $\hat{\beta}_I$ is defined as the solution of equation (3).

Define for each i the $n \times n$ diagonal matrix $A_i = \text{diag}\{a''(\theta_{it})\}$. Under mild regularity conditions we have the following theorem.

THEOREM 1. *The estimator $\hat{\beta}_I$ of β is consistent and $K^{\frac{1}{2}}(\hat{\beta}_I - \beta)$ is asymptotically multivariate Gaussian as $K \to \infty$ with zero mean and covariance matrix V_1 given by*

$$V_I = \lim_{K\to\infty} K\left(\sum_{i=1}^{K} X_i^{\mathrm{T}} \Delta_i A_i \Delta_i X_i\right)^{-1}\left(\sum_{i=1}^{K} X_i^{\mathrm{T}} \Delta_i \operatorname{cov}(Y_i) \Delta_i X_i\right)\left(\sum_{i=1}^{K} X_i^{\mathrm{T}} \Delta_i A_i \Delta_i X_i\right)^{-1}$$

$$= \lim_{K\to\infty} K\{H_1(\beta)\}^{-1} H_2(\beta) \{H_1(\beta)\}^{-1}, \tag{4}$$

where the moment calculations for the Y_i's are taken with respect to the true underlying model.

The proof of the theorem is straightforward and is omitted. The variance of $\hat{\beta}_I$ given in Theorem 1 can be consistently estimated by

$$\{H_1(\hat{\beta}_I)\}^{-1}\left(\left[\sum_{i=1}^{K} X_i^{\mathrm{T}} \Delta_i S_i S_i^{\mathrm{T}} \Delta_i X_i\right]_{\hat{\beta}_I}\right)\{H_1(\hat{\beta}_I)\}^{-1}.$$

Note that the estimation of ϕ is unnecessary for estimating V_I even though the latter is a function of ϕ.

The estimator $\hat{\beta}_I$ has several advantages. It is easy to compute with existing software, e.g. GLIM (Baker & Nelder, 1978). Both $\hat{\beta}_I$ and var $(\hat{\beta}_I)$ are consistent given only a correct specification of the regression which is the principal interest. Note that this requires missing data to be missing completely at random in the sense of Rubin (1976). As discussed in § 5, $\hat{\beta}_I$ can be shown to be reasonably efficient for a few simple designs. The principal disadvantage of $\hat{\beta}_I$ is that it may not have high efficiency in cases where the autocorrelation is large. The next section proposes a 'generalized' estimating equation that leads to estimators with higher efficiency.

3. GENERALIZED ESTIMATING EQUATIONS

3·1. *General*

In this section, we present a class of estimating equations which take the correlation into account to increase efficiency. The resulting estimators of β remain consistent. In addition, consistent variance estimates are available under the weak assumption that a weighted average of the estimated correlation matrices converges to a fixed matrix.

To begin, let $R(\alpha)$ be a $n \times n$ symmetric matrix which fulfills the requirement of being a correlation matrix, and let α be an $s \times 1$ vector which fully characterizes $R(\alpha)$. We refer to $R(\alpha)$ as a 'working' correlation matrix.

Define

$$V_i = A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}}/\phi, \tag{5}$$

which will be equal to cov (Y_i) if $R(\alpha)$ is indeed the true correlation matrix for the Y_i's.

We define the general estimating equations to be

$$\sum_{i=1}^{K} D_i^{\mathrm{T}} V_i^{-1} S_i = 0, \tag{6}$$

where $D_i = d\{a_i'(\theta)\}/d\beta = A_i \Delta_i X_i$. Two remarks are worth mentioning. First, equation (6) reduces to the independence equations in § 2 if we specify $R(\alpha)$ as the identity matrix.

Second, for each i, $U_i(\beta, \alpha) = D_i^{\mathrm{T}} V_i^{-1} S_i$ is similar to the function derived from the quasi-likelihood approach advocated by Wedderburn (1974) and McCullagh (1983) except that the V_i's here are not only a function of β but of α as well. Equation (6) can be reexpressed as a function of β alone by first replacing α in (5) and (6) by $\hat{\alpha}(Y, \beta, \phi)$, a $K^{\frac{1}{2}}$-consistent estimator of α when β and ϕ are known, that is $\hat{\alpha}$ for which $K^{\frac{1}{2}}(\hat{\alpha}-\alpha) = O_p(1)$. Except for particular choices of R and $\hat{\alpha}$, the scale parameter ϕ will generally remain in (6). To complete the process, we replace ϕ by $\hat{\phi}(Y, \beta)$, a $K^{\frac{1}{2}}$-consistent estimator when β is known. Consequently, (6) has the form

$$\sum_{i=1}^{K} U_i[\beta, \hat{\alpha}\{\beta, \hat{\phi}(\beta)\}] = 0, \tag{7}$$

and $\hat{\beta}_G$ is defined to be the solution of equation (7). The next theorem states the large-sample property for $\hat{\beta}_G$.

THEOREM 2. *Under mild regularity conditions and given that:*

(i) $\hat{\alpha}$ *is* $K^{\frac{1}{2}}$*-consistent given* β *and* ϕ*;*

(ii) $\hat{\phi}$ *is* $K^{\frac{1}{2}}$*-consistent given* β*; and*

(iii) $|\partial\hat{\alpha}(\beta, \phi)/\partial\phi| \leqslant H(Y, \beta)$ *which is* $O_p(1)$,

then $K^{\frac{1}{2}}(\hat{\beta}_G - \beta)$ *is asymptotically multivariate Gaussian with zero mean and covariance matrix* V_G *given by*

$$V_G = \lim_{K\to\infty} K\left(\sum_{i=1}^{K} D_i^{\mathrm{T}} V_i^{-1} D_i\right)^{-1} \left\{\sum_{i=1}^{K} D_i^{\mathrm{T}} V_i^{-1} \operatorname{cov}(Y_i)\, V_i^{-1} D_i\right\} \left(\sum_{i=1}^{K} D_i^{\mathrm{T}} V_i^{-1} D_i\right)^{-1}.$$

A sketch of the proof is given in the Appendix. The variance estimate $\hat{V}_G$ of $\hat{\beta}_G$ can be obtained by replacing $\operatorname{cov}(Y_i)$ by $S_i S_i^{\mathrm{T}}$ and β, ϕ, α by their estimates in the expression V_G. As in the independence case, the consistency of $\hat{\beta}_G$ and $\hat{V}_G$ depends only on the correct specification of the mean, not on the correct choice of R. This again requires that missing observations be missing completely at random (Rubin, 1976). Note that the asymptotic variance of $\hat{\beta}_G$ does not depend on choice of estimator for α and ϕ among those that are $K^{\frac{1}{2}}$-consistent. Analogous results are known for the Gaussian data case and in quasi-likelihood where the variance of the regression parameters does not depend on the choice of estimator of ϕ. In our problem, where the likelihood is not fully specified, the result follows from choosing estimating equations for β in which an individual's contribution, U_i, is a product of two terms: the first involving α but not the data, and the second independent of α and with expectation zero. Then $\Sigma E(\partial U_i/\partial\alpha)$ is $o_p(K)$ and $\operatorname{var}(\hat{\beta}_G)$ does not depend on $\hat{\alpha}$ or $\hat{\phi}$ as can be seen from the discussion in the Appendix.

3·2. *Connection with the Gauss–Newton method*

To compute $\hat{\beta}_G$, we iterate between a modified Fisher scoring for β and moment estimation of α and ϕ. Given current estimates $\hat{\alpha}$ and $\hat{\phi}$ of the nuisance parameters, we suggest the following modified iterative procedure for β:

$$\hat{\beta}_{j+1} = \hat{\beta}_j - \left\{\sum_{i=1}^{K} D_i^{\mathrm{T}}(\hat{\beta}_j)\, \tilde{V}_i^{-1}(\hat{\beta}_j)\, D_j(\hat{\beta}_j)\right\}^{-1} \left\{\sum_{i=1}^{K} D_i^{\mathrm{T}}(\hat{\beta}_j)\, \tilde{V}_i^{-1}(\hat{\beta}_j)\, S_i(\hat{\beta}_j)\right\}, \tag{8}$$

where $\tilde{V}_i(\beta) = V_i[\beta, \hat{\alpha}\{\beta, \hat{\phi}(\beta)\}]$. This procedure can be viewed as a modification of Fisher's scoring method in that the limiting value of the expectation of the derivative of $\Sigma U_i[\beta, \hat{\alpha}\{\beta, \hat{\phi}(\beta)\}]$ is used for correction.

Now, define $D = (D_1^{\mathrm{T}}, \ldots, D_K^{\mathrm{T}})^{\mathrm{T}}$, $S = (S_1^{\mathrm{T}}, \ldots, S_K^{\mathrm{T}})$ and let $\tilde{V}$ be a $nK \times nK$ block diagonal matrix with $\tilde{V}_i$'s as the diagonal elements. Define the modified dependent variable

$$Z = D\beta - S,$$

and then the iterative procedure (8) for calculating $\hat{\beta}_G$ is equivalent to performing an iteratively reweighted linear regression of Z on D with weight $\tilde{V}^{-1}$.

3·3. *Estimators of α and ϕ*

At a given iteration the correlation parameters α and scale parameter ϕ can be estimated from the current Pearson residuals defined by

$$\hat{r}_{it} = \{y_{it} - a'(\hat{\theta}_{it})\}/\{a''(\hat{\theta}_{it})\}^{\frac{1}{2}},$$

where $\hat{\theta}_{it}$ depends upon the current value for β. We can estimate ϕ by

$$\hat{\phi}^{-1} = \sum_{i=1}^{K} \sum_{t=1}^{n_i} \hat{r}_{it}^2/(N-p),$$

where $N = \Sigma n_i$. This is the longitudinal analogue of the familiar Pearson statistic (Wedderburn, 1974; McCullagh, 1983). It is easily shown to be $K^{\frac{1}{2}}$-consistent given that the fourth moments of the y_{it}'s are finite. To estimate α consistently, we borrow strength over the K subjects. The specific estimator depends upon the choice of $R(\alpha)$. The general approach is to estimate α by a simple function of

$$\hat{R}_{uv} = \sum_{i=1}^{K} \hat{r}_{iu}\hat{r}_{iv}/(N-p).$$

Specific estimators are given in the next section.

Alternative estimators of ϕ such as one based upon the log likelihood described by McCullagh & Nelder (1983, p. 83) are available. Because we do not specify the entire joint distribution of Y_i, the analogous estimators for α are not available. Note, however, that the asymptotic distribution of $\hat{\beta}_G$ does not depend on the specific choice of α and ϕ among those that are $K^{\frac{1}{2}}$-consistent. The finite sample performance of $\hat{\beta}_G$ for a variety of α, ϕ estimators requires further study.

4. Examples

In this section several specific choices of $R(\alpha)$ are discussed. Each leads to a distinct analysis. The number of nuisance parameters and the estimator of α vary from case to case.

Example 1. Let $R(\alpha)$ be R_0, any given correlation matrix. When $R_0 = I$, the identity matrix, we obtain the independence estimating equation. However for any R_0, $\hat{\beta}_G$ and $\hat{V}_G$ will be consistent. Obviously, choosing R_0 closer to the true correlation gives increased efficiency. Note that for any specified R_0, no knowledge on ϕ is required in estimating β and $\mathrm{var}\,(\hat{\beta}_G)$.

Example 2. Let $\alpha = (\alpha_1, \ldots, \alpha_{n-1})^{\mathrm{T}}$, where $\alpha_t = \mathrm{corr}\,(Y_{it}, Y_{i,t+1})$ for $t = 1, \ldots, n-1$. A natural estimator of α_t, given β and ϕ, is

$$\hat{\alpha}_t = \phi \sum_{i=1}^{K} \hat{r}_{it}\hat{r}_{i,t+1}/(K-p).$$

Now let $R(\alpha)$ be tridiagonal with $R_{t,t+1} = \alpha_t$. This is equivalent to the one-dependent model. An estimator of ϕ is unnecessary for calculating $\hat{\beta}_G$ and $\hat{V}_G$ when the α_t's above are used since the ϕ which appears in the formula for $\hat{\alpha}_t$ cancels in the calculation of V_i. As a special case, we can let $s = 1$ and $\alpha_t = \alpha$ $(t = 1, \ldots, n-1)$. Then the common α can be estimated by

$$\hat{\alpha} = \sum_{t=1}^{n-1} \hat{\alpha}_t/(n-1).$$

An extension to m-dependence is straightforward.

Example 3. Let $s = 1$ and assume that $\text{corr}(y_{it}, y_{it'}) = \alpha$ for all $t \neq t'$. This is the exchangeable correlation structure obtained from a random effects model with a random level for each subject, e.g. Laird & Ware (1982). Given ϕ, α can be estimated by

$$\hat{\alpha} = \phi \sum_{i=1}^{K} \sum_{t>t'} \hat{r}_{it}\hat{r}_{it'} \Big/ \left\{ \sum_{i=1}^{K} \tfrac{1}{2} n_i(n_i-1) - p \right\}.$$

As in Examples 1 and 2, ϕ need not be estimated to obtain $\hat{\beta}_G$ and V_G. Note that an arbitrary number of observations and observation times for each subject are possible with this assumption.

Example 4. Let $\text{corr}(y_{it}, y_{it'}) = \alpha^{|t-t'|}$. For y_{it} Gaussian, this is the correlation structure of the continuous time analogue of the first-order autoregressive process, AR-1 (Feller, 1971, p. 89). Since under this model, $E(\hat{r}_{it}\hat{r}_{it'}) \doteq \alpha^{|t-t'|}$, we can estimate α by the slope from the regression of $\log(\hat{r}_{it}\hat{r}_{it'})$ on $\log(|t-t'|)$. Note that an arbitrary number and spacing of observations can be accommodated with this working model. But $\hat{\phi}$ must be calculated in the determination of $\hat{\beta}_G$ and $\hat{V}_G$.

Example 5. Let $R(\alpha)$ be totally unspecified, that is $s = \frac{1}{2}n(n-1)$. Now R can be estimated by

$$\frac{\phi}{K} \sum_{i=1}^{K} A_i^{-\frac{1}{2}} S_i S_i^{\mathrm{T}} A_i^{-\frac{1}{2}}. \tag{9}$$

Note that for this case, equations (6) and (9) together give the actual likelihood equations if the Y_i's follow a multivariate Gaussian distribution. Further, the asymptotic covariance, V_G, reduces to

$$\lim_{K\to\infty} \left\{ \sum_{i=1}^{K} D_i^{\mathrm{T}} \,\text{cov}^{-1}(Y_i)\, D_i/K \right\}^{-1}$$

since R is the true correlation matrix. Again, no estimation of ϕ is required to obtain $\hat{\beta}_G$. However, this assumption is useful only with a small number of observation times.

5. EFFICIENCY CONSIDERATIONS

In this section, we consider two very simple data configurations and ask the following questions: (i) how much more efficient is $\hat{\beta}_G$ than $\hat{\beta}_I$; and (ii) how do $\hat{\beta}_G$ and $\hat{\beta}_I$ compare to the maximum likelihood estimator when further distributional assumptions on the Y_i's are made? To address the first question, consider the generalized linear model with natural link so that

$$\theta_{it} = x_{it}\beta \quad (t = 1, \ldots, 10).$$

We assume that each $X_i = (x_{i1}, \ldots, x_{i,10})'$ is generated from a distribution with mean $(0{\cdot}1, 0{\cdot}2, \ldots, 1{\cdot}0)'$ and finite covariance. Table 1 then gives the asymptotic relative efficiency of $\hat{\beta}_I$ and $\hat{\beta}_G$'s for three distinct correlation assumptions to the generalized estimator in which the correlation matrix is correctly specified. The correlation structures are one-dependent, exchangeable and first-order autoregressive, Examples 2, 3 and 4. The upper and lower entries are for $\alpha = 0{\cdot}3$ and $0{\cdot}7$ respectively.

Table 1. *Asymptotic relative efficiency of $\hat{\beta}_I$ and $\hat{\beta}_G$ to generalized estimator with correlation matrix correctly specified for $\eta_{it} = \beta_0 + \beta_1 t/10$. Here, $\beta_0 = \beta_1 = 1$, $n_i = 10$. For upper entry $\alpha = 0{\cdot}3$; lower entry $\alpha = 0{\cdot}7$*

	Working R			
True R	Independence	1-dependence	Exchangeable	AR-1
1-Dependence	0·97	1·0	0·97	0·99
	0·74	1·0	0·74	0·81
Exchangeable	0·99	0·95	1·0	0·95
	0·99	0·23	1·0	0·72
AR-1	0·97	0·99	0·97	1·0
	0·88	0·75	0·88	1·0

There is little difference between $\hat{\beta}_I$ and the $\hat{\beta}_G$'s when the true correlation is moderate, 0·3 say. However, lower entries of Column 1 indicate that substantial improvement can be made by correctly specifying the correlation matrix when α is large. The efficiency of $\hat{\beta}_I$ relative to the $\hat{\beta}_G$ using the correct correlation matrix is lowest, 0·74, when R has the one-dependent form and highest, 0·99, when R has the exchangeable pattern. That $\hat{\beta}_I$ is efficient relative to $\hat{\beta}_G$ in the latter case is because $n_i = 10$ for all i so that the extra-binomial variation introduced by the exchangeable correlation is the same for all subjects and no misweighting occurs by ignoring it. If instead, we assume that n_i takes values from 1 to 8 with equal probability, the relative efficiency of $\hat{\beta}_I$ drops to 0·82. Note that the results in Table 1 hold regardless of the underlying marginal distribution.

To address the second question, we consider a two-sample configuration with binary outcomes. Subjects are in two groups, with marginal expectations satisfying $\text{logit}\{E(y_{it})\} = \beta_0 + \beta_1 x_i$, where $x_i = 0$ for Group 0, and 1 for Group 1. The repeated observations are assumed to come from a Markov chain of order 1 with first lag autocorrelation α. In Table 2, we compare the asymptotic relative efficiencies of $\hat{\beta}_I$ and

Table 2. *Asymptotic relative efficiency of $\hat{\beta}_I$ and $\hat{\beta}_G$ assuming* AR 1 *correlation structure to the maximum likelihood estimate for first-order Markov chain with $\theta_{it} = \beta_0 + \beta_1 x_i, x_i = 0$ for Group 0, $x_i = 1$ for Group 1. Here $\beta_0 = 0, \beta = 1$, and for upper entry $n_i = 10$, lower entry $n_i = 1, \ldots, 8$ with equal probabilities*

	Correlation, α						
	0·0	0·1	0·2	0·3	0·5	0·7	0·9
$\hat{\beta}_I$	1·0	1·0	0·99	0·97	0·94	0·91	0·92
	1·0	1·0	0·98	0·96	0·92	0·86	0·81
$\hat{\beta}_G$ (AR 1)	1·0	1·0	0·99	0·99	0·98	0·97	0·98
	1·0	1·0	0·99	0·99	0·98	0·98	0·99

$\hat{\beta}_G$ using the AR–1 correlation structure, Example 4, to the maximum likelihood estimator. For the upper entry, $n_i = 10$ for all i; for the lower, $n_i = 1$ to 8 with equal probability. The results indicate that both $\hat{\beta}_I$ and $\hat{\beta}_G$ are highly efficient for smaller α. As α increases, $\hat{\beta}_G$ retains nearly full efficiency while $\hat{\beta}_I$ does not. The contrast between $\hat{\beta}_I$ and $\hat{\beta}_G$ is strongest for the unequal sample size case.

6. DISCUSSION

The analysis of non-Gaussian longitudinal data is difficult partly because few models for the joint distribution of the repeated observations for a subject are available. On the other hand, longitudinal data offer the advantage that data from distinct subjects are independent. The methods we propose avoid the need for multivariate distributions by only assuming a functional form for the marginal distribution at each time. The covariance structure across time is treated as a nuisance. We rely, however, on the independence across subjects to estimate consistently the variance of the proposed estimators even when the assumed correlation is incorrect, as we expect it often will be.

Modelling the marginal expectation and treating the correlation as a nuisance may be less appropriate when the time course of the outcome for each subject, e.g. growth, is of primary interest or when the correlation itself has scientific relevance. The random effects model for binary data discussed by Stiratelli, Laird & Ware (1984) can be extended to the generalized linear model family and is more appropriate for the study of growth. When the time dependence is central, models for the conditional distribution of y_t given $y_{t-1}, y_{t-2}, \ldots, y_1$ may be more appropriate. Cox (1970, p. 72) has proposed such a model for binary outcomes. Korn & Whittemore (1979) have applied this model to air pollution data.

The examples in §4 provide several alternative methods for analysing longitudinal data sets. The method in Example 1, which includes the independence estimating equation as a special case, requires the fewest assumptions. Only the regression specification must be correct to obtain consistent estimates of β and var $(\hat{\beta})$. In §5, the independence estimator was shown to have high efficiency when the correlation is moderate in a simple situation with binary outcomes, Table 2. We believe that it may be less efficient in more realistic situations with more heterogeneity among both the X_i's and n_i's. Further study is needed.

Among the remaining methods implied by the generalized estimating equation, allowing R to have $\frac{1}{2}n(n-1)$ parameters, Example 5, gives the most efficient estimator. This approach, however, is only useful when there are few observation times. The remaining estimators will be as efficient only if the true correlation matrix can be expressed in terms of the chosen $R(\alpha)$ for some α. In particular, all generalized estimating equation estimators will be efficient if observations for a subject are independent. Note that each estimator and its variance will be consistent as long as α and ϕ can be estimated consistently for any correlation.

Missing data are common in some longitudinal studies. For $\hat{\beta}_G$ and $\hat{V}_G$ to be consistent even when R is misspecified, we require that data be missing completely at random (Rubin, 1976). That is, whether an observation is missing cannot depend on previous outcomes. Intuitively, we should not expect to handle complicated missing value patterns unless our working model is correct. When R is the true correlation, the missing completely at random assumption can be unnecessary. For Gaussian outcomes, the missing data pattern can depend arbitrarily on past observations and consistency is

retained. For binary outcomes, the pattern can depend on any single previous outcome.

If the elements of R are proportional to those of α, then the scale parameter, ϕ, does not have to be determined as a step in solving the general estimating equation. This was the case for all examples above except 4. Note that ϕ also is eliminated from the estimation of β in quasi-likelihood methods (Wedderburn, 1974). In addition, the variance of $\hat{\beta}_G$ does not depend on the choice of estimator of the nuisance parameters, α and ϕ among those that are $K^{\frac{1}{2}}$-consistent. This is also the case in quasi-likelihood where the only nuisance parameter is ϕ. The estimating equations described in this paper can be thought of as an extension of quasi-likelihood to the case where the second moment cannot be fully specified in terms of the expectation but rather additional correlation parameters must be estimated. It is the independence across subjects that allows us to consistently estimate these nuisance parameters where this could not be done otherwise.

Acknowledgements

We thank the referee and Professor Nan Laird for helpful comments.

Appendix

Proof of Theorem 2

Write $\alpha^*(\beta) = \hat{\alpha}\{\beta, \hat{\phi}(\beta)\}$ and under some regularity conditions $K^{\frac{1}{2}}(\hat{\beta}_G - \beta)$ can be approximated by

$$\left[\sum_{i=1}^{K} -\frac{\delta}{\delta\beta} U_i\{\beta, \alpha^*(\beta)\}/K\right]^{-1}\left[\sum_{i=1}^{K} U_i\{\beta, \alpha^*(\beta)\}/K^{\frac{1}{2}}\right],$$

where

$$\begin{aligned}\delta U_i\{\beta, \alpha^*(\beta)\}/\delta\beta &= \partial U_i\{\beta, \alpha^*(\beta)\}/\partial\beta + [\partial U_i\{\beta, \alpha^*(\beta)\}/\partial\alpha^*]\{\partial\alpha^*(\beta)/\partial\beta\} \\ &= A_i + B_i C. \end{aligned} \tag{A1}$$

Let β be fixed and Taylor expansion gives

$$\begin{aligned}\frac{\Sigma U_i\{\beta, \alpha^*(\beta)\}}{K^{\frac{1}{2}}} &= \frac{\Sigma U_i(\beta, \alpha)}{K^{\frac{1}{2}}} + \frac{\Sigma \partial/\partial\alpha\, U_i(\beta, \alpha)}{K} K^{\frac{1}{2}}(\alpha^* - \alpha) + o_p(1) \\ &= A^* + B^* C^* + o_p(1), \end{aligned} \tag{A2}$$

where the sums are over $i = 1, \ldots, K$. Now, $B^* = o_p(1)$, since $\partial U_i(\beta, \alpha)/\partial\alpha$ are linear functions of S_i's whose means are zero, and conditions (i) to (iii) give

$$\begin{aligned} C^* &= K^{\frac{1}{2}}[\hat{\alpha}\{\beta, \hat{\phi}(\beta)\} - \hat{\alpha}(\beta, \phi) + \hat{\alpha}(\beta, \phi) - \alpha] \\ &= K^{\frac{1}{2}}\left\{\frac{\partial\hat{\alpha}}{\partial\phi}(\beta, \phi^*)(\hat{\phi} - \phi) + \hat{\alpha}(\beta, \phi) - \alpha\right\} = O_p(1). \end{aligned}$$

Consequently, $\Sigma U_i\{\beta, \alpha^*(\beta)\}/K^{\frac{1}{2}}$ is asymptotically equivalent to A^* whose asymptotic distribution is multivariate Gaussian with zero mean and covariance matrix

$$\lim_{K\to\infty}\left\{\sum_{i=1}^{K} D_i^{\mathrm{T}} V_i^{-1} \operatorname{cov}(Y_i) V_i^{-1} D_i/K\right\}.$$

Finally, it is easy to see that $\Sigma B_i = o_p(K)$, $C = O_p(1)$ and that $\Sigma A_i/K$ converges as $K \to \infty$ to $-\Sigma D_i^{\mathrm{T}} V_i^{-1} D_i/K$. This completes the proof.

REFERENCES

BAKER, R. J. & NELDER, J. A. (1978). *The GLIM System, Release* 3. *Generalized Linear Interactive Modelling*. Oxford: Numerical Algorithms Group.

COX, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

FELLER, W. (1971). *An Introduction to Probability Theory*, **2**, 2nd ed. New York: Wiley.

JORGENSEN, B. (1983). Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika* **70**, 19–28.

KOCH, G. C., LANDIS, J. R., FREEMAN, J. L., FREEMAN, D. H. & LEHMAN, R. G. (1977). A general methodology for the analysis of repeated measurements of categorical data. *Biometrics* **33**, 133–58.

KORN, E. L. & WHITTEMORE, A. S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics* **35**, 795–802.

LAIRD, N. M. & WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–74.

MCCULLAGH, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11**, 59–67.

MCCULLAGH, P. & NELDER, J. A. (1983). *Generalized Linear Models*. London: Chapman and Hall.

MORTON, R. (1981). Efficiency of estimating equations and the use of pivots. *Biometrika* **68**, 227–33.

OCHI, Y. & PRENTICE, R. L. (1984). Likelihood inference in correlated probit regression. *Biometrika* **71**, 531–43.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63**, 81–92.

STIRATELLI, R., LAIRD, N. & WARE, J. (1984). Random effects models for serial observations with binary responses. *Biometrics* **40**, 961–71.

WARE, J. H. (1985). Linear models for the analysis of longitudinal studies. *Am. Statistician* **39**, 95–101.

WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61**, 439–47.

ZEGER, S. L., LIANG, K. Y. & SELF, S. G. (1985). The analysis of binary longitudinal data with time independent covariates. *Biometrika* **72**, 31–8.

[*Received January* 1985. *Revised October* 1985]