Longitudinal Data Analysis for Discrete and Continuous Outcomes

Scott L. Zeger; Kung-Yee Liang

Biometrics, Vol. 42, No. 1. (Mar., 1986), pp. 121-130.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28198603%2942%3A1%3C121%3ALDAFDA%3E2.0.CO%3B2-E

Biometrics is currently published by International Biometric Society.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ibs.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org
Tue May 8 03:45:19 2007

BIOMETRICS 42, 121–130
March 1986

Longitudinal Data Analysis for Discrete and Continuous Outcomes

Scott L. Zeger and Kung-Yee Liang

Department of Biostatistics, Johns Hopkins University,
School of Hygiene and Public Health,
615 N. Wolfe Street, Baltimore, Maryland 21205, U.S.A.

SUMMARY

Longitudinal data sets are comprised of repeated observations of an outcome and a set of covariates for each of many subjects. One objective of statistical analysis is to describe the marginal expectation of the outcome variable as a function of the covariates while accounting for the correlation among the repeated observations for a given subject. This paper proposes a unifying approach to such analysis for a variety of discrete and continuous outcomes.

A class of generalized estimating equations (GEEs) for the regression parameters is proposed. The equations are extensions of those used in quasi-likelihood (Wedderburn, 1974, *Biometrika* **61,** 439–447) methods. The GEEs have solutions which are consistent and asymptotically Gaussian even when the time dependence is misspecified as we often expect. A consistent variance estimate is presented. We illustrate the use of the GEE approach with longitudinal data from a study of the effect of mothers' stress on children's morbidity.

1. Introduction

Longitudinal data sets are comprised of repeated observations of an outcome variable and a set of covariates for each of many subjects. For example, in a study of the impact of mothers' stress on children's morbidity, the presence (1) or absence (0) of illness in children as well as the level of mothers' stress were recorded daily for a sample of children. One possible objective in analyzing longitudinal data sets is to describe the marginal expectation of the outcome, here the probability of illness, as a function of the predictor variables. Because repeated observations are made on each subject, correlation is anticipated among a subject's measurements. It must be accounted for to obtain a correct statistical analysis.

When the outcome variable is approximately Gaussian, a large class of linear models is available for analysis. Rao (1965), Grizzle and Allen (1969), and Hui (1984) have discussed methods based on fitting growth curves to the repeated observations for each subject. Here, each subject's data are modelled as a simple function of time and the dependence of the resulting coefficients on the covariates is then assessed. Fearn (1975) discussed a Bayesian approach to growth curve modelling. Harville (1977) and Laird and Ware (1982) developed random-effects models in which repeated observations for a subject are assumed to share a common random component. Laird and Ware's (1982) discussion is more general, including both growth models and random-effects models as special cases. Azzalini (1984) discussed models in which autoregressive error structure was assumed. Here, the autocorrelation decreases as a geometric function of the time between two observations. Ware (1985) has presented an overview of linear models for Gaussian longitudinal data.

This is Department of Biostatistics paper #568, Johns Hopkins University.

Key words: Discrete data; Generalized estimating equations; Generalized linear models; Longitudinal data; Quasi-likelihood; Repeated measures.

Fewer techniques have been available when the outcome is not approximately Gaussian. Random-effects models for binary outcomes are the exception. Stiratelli, Laird, and Ware (1984) and Anderson and Aitkin (1985) have developed a logistic model; Ochi and Prentice (1984), a probit model; and Koch et al. (1977), log-linear models. Of these, only the logistic random-effects model allows for time-dependent covariates. Cox (1970) alternatively suggested a logistic regression model for the conditional expectation of binary data given previous observations. This model corresponds to a Markov chain in which the transition probabilities depend on the covariates. Korn and Whittemore (1979) have used this method in a study of the health effects of air pollution.

Zeger, Liang, and Self (1985) proposed a class of Markov "working" models for binary longitudinal data. They showed that as the number of subjects increased, the regression coefficients obtained from the working likelihood analysis were consistent and had consistent variance estimates under weak assumptions about the actual time dependence. However, their models, which explicitly included parameters for the time dependence, could be used only with time-independent covariates.

The limited number of models for non-Gaussian longitudinal data is partly due to the lack of a rich class of distributions, such as the multivariate Gaussian, for the repeated observations for each subject. Hence, likelihood analyses have not been developed except in the particular cases mentioned above. Even in the binary case where likelihood analysis is possible, computation is difficult (Stiratelli et al., 1984).

In the regression context with a single observation for each subject, generalized linear models and quasi-likelihood theory (McCullagh and Nelder, 1983; Wedderburn, 1974) have extended linear models from the Gaussian case to a broad class of outcomes. In the quasi-likelihood approach, a known transformation of the marginal expectation of the outcome is assumed to be a linear function of the covariates. Instead of specifying the distribution of the dependent variable, we assume its variance is a known function of its expectation. For example, with binary outcomes we might assume that the logit of the probability of response, p, depends linearly on the covariates. The variance is just $p(1 - p)$. This partial specification of the outcome distribution leads to simple techniques for regression analyses of Gaussian, gamma, Poisson, binomial, categorical, and ordinal data (McCullagh and Nelder, 1983).

In this paper, we propose a methodology for discrete and continuous longitudinal data that uses the quasi-likelihood approach. We turn to quasi-likelihood because of the sparseness of multivariate distributions for non-Gaussian data. We specify that a known function of the marginal expectation of the dependent variate is a linear function of the covariates, and assume that the variance is a known function of the mean. In addition, we specify a "working" correlation matrix for the observations for each subject. This set-up leads to generalized estimating equations (GEEs) which give consistent estimators of the regression coefficients and of their variances under weak assumptions about the actual correlation among a subject's observations.

Liang and Zeger (1986) recently derived the GEEs discussed here from a different and slightly more limited context. They assumed that the marginal distribution of the dependent variable followed a generalized linear model (McCullagh and Nelder, 1983). They proposed a "working" model in which repeated observations for a subject were assumed to be independent. They then generalized this "independence working model" to explicitly account for correlation, giving the GEE.

The broad objectives of this paper are not to give details about quasi-likelihood and its extension to longitudinal data nor to give technical results about the GEEs. Rather, our intention is to motivate what we consider to be a widely applicable methodology for longitudinal data, summarize its advantages and disadvantages, and illustrate its use. Section 2 briefly describes quasi-likelihood. Section 3 applies the quasi-likelihood approach to

longitudinal data giving the generalized estimating equations (GEE). Section 4 illustrates the methodology with the stress–morbidity data. The final section discusses problems which arise in using GEEs and mentions extensions to other situations.

2. Quasi-Likelihood

This section briefly describes the aspects of quasi-likelihood theory used in our development of the GEE approach to longitudinal data analysis. Quasi-likelihood was first proposed by Wedderburn (1974) and later examined extensively by McCullagh (1983). It is a methodology for regression that requires few assumptions about the distribution of the dependent variable and hence can be used with a variety of outcomes. In likelihood analysis, we must specify the actual form of the distribution. In quasi-likelihood, we specify only the relationships between the outcome mean and covariates and between the mean and variance. This extension is important to our problem in that, except for nearly Gaussian outcomes, there are few choices for the joint distribution of the repeated values for each subject. By adopting a quasi-likelihood approach and specifying only the mean–covariance structure, we can develop methods that are applicable to several types of outcome variables.

To establish notation useful in the next section as well, consider the observations $(y_{ij}, \mathbf{x}_{ij})$ for times t_{ij}, $j = 1, \ldots, n_i$ and subjects $i = 1, \ldots, K$. Here y_{ij} is the outcome variable and $\mathbf{x}_{ij}$ is a $p \times 1$ vector of covariates. Let $\mathbf{y}_i$ be the $n_i \times 1$ vector $(y_{i1}, \ldots, y_{in_i})'$ and $\mathbf{x}_i$ be the $n_i \times p$ matrix $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})'$ for the ith subject. Quasi-likelihood has previously been applied to the regression context where $n_i = 1$ for all i. Hence, in discussing these results in this section, we drop the subscript j and treat each subject's data as a scalar.

Define μ_i to be the expectation of y_i and suppose that

$$\mu_i = h(\mathbf{x}_i\boldsymbol{\beta}) \tag{1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters. The inverse of h is referred to as the "link" function (McCullagh and Nelder, 1983). In quasi-likelihood, the variance, v_i, of y_i is expressed as a known function, g, of the expectation, μ_i, i.e.,

$$v_i = g(\mu_i)/\phi \tag{2}$$

where ϕ is a scale parameter. The focus of quasi-likelihood is on methods for inference about $\boldsymbol{\beta}$. Hence, ϕ is treated as a nuisance parameter.

The quasi-likelihood estimator is the solution of the score-like equation system

$$S_k(\boldsymbol{\beta}) = \sum_{i=1}^{K} \frac{\partial \mu_i}{\partial \beta_k} v_i^{-1}(y_i - \mu_i) = 0, \quad k = 1, \ldots, p. \tag{3}$$

Equations (3) are in fact score equations for $\boldsymbol{\beta}$ when y_i has distribution from the exponential family. Their solution can be obtained by an iteratively reweighted least squares. The resulting estimator is asymptotically Gaussian under mild regularity conditions (McCullagh, 1983). It also possesses a Gauss–Markov-like optimality in that it is asymptotically the minimum variance estimator among those with linear influence function. Wedderburn (1974) and McCullagh (1983) provide details about quasi-likelihood in the regression context.

3. Generalized Estimating Equations (GEEs) for Longitudinal Data

To apply the quasi-likelihood approach to the analysis of longitudinal data, we must consider the mean and covariance of the vector of responses, $\mathbf{y}_i$, for the ith subject. We proceed as in Section 2 but in addition let $\mathbf{R}_i(\boldsymbol{\alpha})$ be the $n_i \times n_i$ "working" correlation matrix for each $\mathbf{y}_i$. Note that the observation times and correlation matrix can differ from subject to subject. $\mathbf{R}_i(\boldsymbol{\alpha})$, however, is assumed to be fully specified by the $s \times 1$ vector of unknown

parameters, α, which is the same for all subjects. Then following the quasi-likelihood approach, the working covariance matrix for $\mathbf{y}_i$ is given by

$$\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i(\alpha)\mathbf{A}_i^{1/2}/\phi, \tag{4}$$

where $\mathbf{A}_i$ is an $n_i \times n_i$ diagonal matrix with $g(\mu_{ij})$ as the jth diagonal element. We refer to $\mathbf{R}_i(\alpha)$ as a "working" correlation matrix because we do not expect it to be correctly specified. We would like estimators that are consistent and have consistent variance estimates even when $\mathbf{R}_i(\alpha)$ is incorrect.

Our extension of equations (3) to the longitudinal data case is given by

$$\sum_{i=1}^{K} \mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{S}_i = \mathbf{0}. \tag{5}$$

Here $\mathbf{S}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$ with $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in})'$ and $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\beta$. Equations (5) reduce to the quasi-likelihood equations (3) when $n_i = 1$ for all i. More generally, $\mathbf{U}_i(\beta, \alpha) = \mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{S}_i$ is equivalent to the estimating function suggested by Wedderburn (1974) except that the $\mathbf{V}_i$'s here are functions of α as well as β. Equations (5) are designed to guarantee consistency of the regression coefficients when the link function is correctly specified under minimal assumptions about the time dependence. Note that $\mathbf{D}_i'\mathbf{V}_i^{-1}$ does not depend on the y's so that equations (5) converge to $\mathbf{0}$ and hence have consistent roots as long as $\mathrm{E}\mathbf{S}_i = \mathbf{0}$. Finally, note that for Gaussian outcomes, equations (5) are the score equations for β.

While the estimating equations now depend on α as well as β, they can be reexpressed as a function of β alone by first replacing α in equations (4) and (5) by a $K^{1/2}$-consistent estimator, $\hat{\alpha}(\mathbf{Y}, \beta, \phi)$, and then replacing ϕ in $\hat{\alpha}$ by a $K^{1/2}$-consistent estimator, $\hat{\phi}(\mathbf{y}, \beta)$. Consequently, for any given $\mathbf{R}_i(\alpha)$, the estimate, $\hat{\beta}_R$, of β is defined as the solution of

$$\sum_{i=1}^{K} \mathbf{U}_i\{\beta, \hat{\alpha}[\beta, \hat{\phi}(\beta)]\} = \mathbf{0}. \tag{6}$$

Under mild regularity conditions, Liang and Zeger (1986, Theorem 2) show that as $K \to \infty$, $\hat{\beta}_R$ is a consistent estimator of β and that $K^{1/2}(\hat{\beta}_R - \beta)$ is asymptotically multivariate Gaussian with covariance matrix $\mathbf{V}_R$ given by

$$\begin{aligned}\mathbf{V}_R &= \lim_{K\to\infty} K\left(\sum_{i=1}^{K} \mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i\right)^{-1}\left[\sum_{i=1}^{K} \mathbf{D}_i'\mathbf{V}_i^{-1}\mathrm{cov}(\mathbf{y}_i)\mathbf{V}_i^{-1}\mathbf{D}_i\right]\left(\sum_{i=1}^{K} \mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i\right)^{-1} \\ &= \lim_{K\to\infty} K(\mathbf{V}_1^{-1}\mathbf{V}_0\mathbf{V}_1^{-1}),\end{aligned} \tag{7}$$

where the covariance of $\mathbf{y}_i$ is the actual rather than the assumed covariance. $\mathbf{V}_R$ can be estimated consistently without evaluating $\mathrm{cov}(\mathbf{y}_i)$ directly. This is achieved by simply replacing $\mathrm{cov}(\mathbf{y}_i)$ by $\mathbf{S}_i\mathbf{S}_i'$ and α, β, and ϕ by their estimates in (7). It is interesting to note that the asymptotic covariance matrix of $\hat{\beta}_R$ does not depend on the choice of α and ϕ among those that are $K^{1/2}$-consistent.

To solve the GEE for $\hat{\beta}_R$, we iteratively solve for the regression coefficients and the correlation and scale parameters, α and ϕ. Given an estimate of $\mathbf{R}_i(\alpha)$ and of ϕ, we can calculate an updated estimate of β by iteratively reweighted least squares as described by McCullagh and Nelder (1983). GLIM (Baker and Nelder, 1978) can be used in this step. Given an estimate of β, we calculate standardized residuals, $r_{ij} = (y_{ij} - \hat{\mu}_{ij})/\sqrt{[\hat{\mathbf{V}}_1^{-1}]_{jj}}$, which are used to consistently estimate α and ϕ. These two steps are iterated until convergence. Details on computing $\hat{\beta}_R$ and $\hat{\mathbf{V}}_R$ are provided by Liang and Zeger (1986). As in many quasi-likelihood problems, it is often possible to estimate β without estimating ϕ directly. We require only that the elements of $\mathbf{R}$ be multiples of the parameters, α. This is the case for many choices of practical interest.

A useful feature of the GEE approach is that it is not necessary for the "working" correlation matrix to be correctly specified to obtain a consistent and asymptotically Gaussian estimate, $\hat{\boldsymbol{\beta}}_R$, or for estimating $\mathbf{V}_R$ consistently. We require only that $\boldsymbol{\alpha}$ and ϕ be estimated consistently and that the matrix $\mathbf{V}_1$ on the right-hand side in equation (7) converge when divided by K to a fixed matrix. Hence, it is also not necessary that the observations for all subjects have the same correlation structure. However, this robustness property holds only when there is a diminishing fraction of missing data or when the data are missing completely at random (Rubin, 1976).

There are several choices for the working correlation matrix, $\mathbf{R}_i$. The simplest is to assume $\mathbf{R}_i = \mathbf{I}_{n_i}$, the $n_i \times n_i$ identity matrix, i.e., that repeated observations are uncorrelated. For $\mathbf{y}$'s from the exponential family, the GEEs (7) then reduce to the score equations obtained by assuming repeated observations are independent. The resulting estimator is a generalization of the one proposed by Zeger et al. (1985) for binary outcomes. A second extreme case is applicable when observation times are the same for all subjects so that $\mathbf{R}_i(\boldsymbol{\alpha}) = \mathbf{R}(\boldsymbol{\alpha})$ and $n_i = n$. We may then let $\mathbf{R}(\boldsymbol{\alpha})$ be fully unspecified and estimate the $n(n-1)/2$ correlations. The resulting estimator is asymptotically minimum variance among the class of estimators which satisfy the GEE for different choices of $\mathbf{R}$. Another choice for many longitudinal studies is to let $[\mathbf{R}_i]_{jk} = \alpha$, $j \neq k$. This is the correlation assumed in a random-effects model. Finally, $\mathbf{R}(\boldsymbol{\alpha})$ might be chosen so that

$$[\mathbf{R}_i]_{jk} = \begin{cases} \alpha^{|t_{ij}-t_{ik}|}, & |t_{ij}-t_{ik}| \leq m \\ 0, & |t_{ij}-t_{ik}| > m \end{cases},$$

where t_{ij}, t_{ik} are the jth and kth observation times for the ith subject. This is the correlation structure for a stationary m-dependent process.

Because both $\hat{\boldsymbol{\beta}}_R$ and $\hat{\mathbf{V}}_R$ are robust to the choice of $\mathbf{R}_i$, confidence intervals for $\boldsymbol{\beta}$ and other statistical assessments of the model will be asymptotically correct even when $\mathbf{R}$ is misspecified, as we believe will be the case more often than not. Choosing the working correlation matrix to be close to the actual one, however, increases efficiency as, for example, in the case when the outcomes are Gaussian.

This GEE approach is currently applicable to longitudinal data analyses with univariate outcomes for which the quasi-likelihood formulation is sensible. This includes Gaussian, Poisson, binomial (binary), gamma, and inverse Gaussian variables. In addition, the GEE approach can be extended to include multinomial and ordinal data which are multivariate by nature. By choosing $\mathbf{R}_i$ to be the identity matrix, we can currently use formulations described by McCullagh and Nelder (1983) for multinomial and ordinal variables. Other choices of $\mathbf{R}_i$ need to take into account the multivariate nature of the outcome and will be discussed elsewhere.

4. Example: Mothers' Stress–Children's Morbidity

This section illustrates the GEE approach with data from a study of the association of mothers' stress and children's morbidity. A population of mothers with infants between the ages of 18 months and 5 years were recruited by Professor Cheryl Alexander of the Johns Hopkins University Department of Maternal and Child Health from the Ambulatory Pediatric Health Services Clinic at Baltimore City Hospital, Baltimore, Maryland. They were asked to keep a diary and to record daily whether their child was ill and their own relative stress. In a preliminary interview, time-independent covariates including child's race, household size, mother's marital status, employment status, and general level of stress were determined. We have used the first 9 days of the diaries for 167 women with nearly complete records to illustrate use of the GEE method. Hence, $n_i = n = 9$ and $K = 167$. Note that the outcome and all predictor variables in this analysis are dichotomous. Table 1 summarizes the proportion of positive responses for each variable over all times and subjects.

Table 1
Frequencies for dichotomous variables averaged over all times and children

Variable	Frequency of response 1
Children's illness	.14
Mother's stress, $t-1$	.14
Mother's stress, $t-2$	.15
Mother's stress, $t-3$	.16
Household size	.66
0: 2–3	
1: >3	
Race	.55
0: White	
1: Nonwhite	
Mother's general stress	.55
0: Low	
1: High	
Employment	33
0: Unemployed	
1: Employed	
Marital status	.50
0: Other	
1: Married	

We assumed that the logit of the probability that a child was ill on day t is a linear function of the mother's stress on days $t-1$, $t-2$, and $t-3$, and on the time-independent covariates listed above. Table 2 presents the estimated regression coefficients and the associated t-statistics obtained using four separate choices of the working correlation matrix, $\mathbf{R}_i = \mathbf{R}$. We first assumed that $\mathbf{R} = \mathbf{I}$, i.e., that repeated outcomes for a given child were independent. Mother's stress 3 days prior and household size were found to be strongly associated with the probability of morbidity for the child. Race and mother's general stress level were marginally associated. Mothers who rated themselves as being stressed more than average on all three days ($t-1$, $t-2$, $t-3$) were estimated to have a 2.5-fold increase in the odds of their child being sick over women who were stressed on none of the days. Nonwhites had 1.6 times the odds of illness over whites.

Similar results were obtained when the working correlation matrix, $\mathbf{R}$, was assumed to be either 1-dependent or stationary. In the first case, observations were assumed to be correlated only with those immediately before or after them. The first lag correlation coefficient was estimated to be .34. In the latter case, the correlation was assumed to depend only on the time separating two observations. Here the correlations corresponding to lags 1 to 8 were estimated to be .35, .17, .07, .08, .02, −.07, −.09, .05. Note that the correlation decreases to near 0 after the first or second lag so that the 1-dependent and stationary correlation models should give similar results. This can be seen in Table 2. In addition, the coefficients and t-statistics for these alternative models are similar to those obtained for the independence working model. The coefficients for the time-dependent mother's stress variables decreased somewhat while those for the time-independent covariates changed very little. The effect of assuming a nonindependent correlation structure is to use weighted linear combinations of both the y's and x's for each subject in the GEE. It is therefore sensible that the covariates which vary with time may experience larger changes in their coefficients than do the time-independent covariates for finite samples. Note, however, that the qualitative conclusions are the same for each of these three choices of $\mathbf{R}$. The results obtained here are also qualitatively similar to those reported by Kane (unpublished Ph.D.

Table 2
Estimated regression coefficients and t-statistics for 4 *different choices of* **R**

	Coefficients				*t*-statistics			
	Independence	1-Dependence	Stationary	Exchangeable	Independence	1-Dependence	Stationary	Exchangeable
Intercept	−2.12	−2.05	−2.02	−1.99	−6.99	−6.78	−6.79	−6.35
Mother's stress $t-1$	0.21	0.09	0.12	−0.12	1.14	0.47	0.62	−0.47
Mother's stress $t-2$	0.25	0.23	0.15	−0.06	1.35	1.21	0.79	−0.26
Mother's stress $t-3$	0.44	0.39	0.38	0.22	2.67	2.35	2.47	1.17
Household size	−0.67	−0.69	−0.68	−0.74	−2.92	−3.05	−3.10	−3.14
Race	0.46	0.50	0.51	0.45	2.02	2.17	2.25	1.95
Mother's general stress	0.36	0.32	0.30	0.43	1.55	1.40	1.31	1.83
Employment	−0.23	−0.25	−0.22	−0.22	−0.98	−1.07	−0.96	−0.84
Marital status	0.35	0.35	0.34	0.39	1.54	1.57	1.58	1.70

Table 3
Comparison of t-statistics from GEE method with naive t-statistics obtained by assuming working correlation matrix is correct

	Independence		1-Dependence	
	t-statistic	Naive *t*-statistic	*t*-statistic	Naive *t*-statistic
Intercept	−6.99	−10.15	−6.78	−7.9
Mother's stress $t-1$	1.14	0.99	0.47	0.47
Mother's stress $t-2$	1.35	1.18	1.21	1.19
Mother's stress $t-3$	2.67	2.26	2.35	2.12
Household size	−2.92	−3.95	−3.05	−3.26
Race	2.02	2.66	2.17	2.28
Mother's general stress	1.55	2.13	1.40	1.49
Employment	−0.98	−1.20	−1.07	−1.02
Marital status	1.54	2.00	1.57	1.60

thesis, Johns Hopkins University, 1984), who used logistic regression and included previous **y**'s to account for the time dependence.

For illustration, we have also assumed an exchangeable correlation matrix, i.e., that $R_{jk} = \alpha, j \neq k$. We let $\hat{\alpha} = .35$, the estimated first lag correlation. Hence, the correlation between the first and last observations for a subject was assumed to be .35, rather than its estimated value .05. Our intention is to illustrate the effect of grossly misspecifying the correlation. For a few predictor variables, the resulting coefficients and *t*-statistics in Table 2 can be seen to differ more markedly from their previous values. The largest difference is for the stress variables which are time-dependent. The coefficients for the time-independent predictors, however, are quite similar to those obtained for more appropriate choices of **R**. The differences are likely due to the limited number of women given the number of covariates. Note that the sensitivity of inferences about β to misspecification of **R** is likely also to depend on the degree and pattern of incomplete data, although that was not a problem in this example.

Table 3 presents a comparison, for two choices of **R**, of the *t*-statistics from the GEE analysis against naive *t*-statistics obtained by assuming that the working correlation matrix, **R**, is the true correlation matrix. The naive *t*-statistic for the *j*th covariate is defined $t_j = \hat{\beta}_j / \sqrt{[\hat{\mathbf{V}}_1^{-1}]_{jj}}$. First consider the comparison for the independence working model. For the time-independent covariates, the naive *t*-values are larger than those obtained from the GEE analysis. They therefore tend to overstate the association. This is expected with positive correlation. Interestingly, the naive *t*-statistics for the time-dependent stress covariates are slightly conservative. For the 1-dependent working correlation matrix, note that the GEE and naive *t*-statistics are more similar. This is an indication that the 1-dependent model is more consistent with the observed association.

5. Discussion

Longitudinal data sets in which the outcome variable cannot be transformed to be Gaussian are more difficult to analyze for two reasons. First, simple models for the conditional expectation of the outcome do not imply equally simple models for the marginal expectation, as is the case for Gaussian data. Hence, the analyst must choose to model either the marginal or conditional expectation. Second, likelihood analyses often lead to estimators of the regression coefficients which are consistent only when the time dependence is correctly specified. The generalized estimating equations for β have been designed to

guarantee consistency of the regression coefficient estimates under minimal assumptions about the time dependence. This approach is sensible when the regression equation for the marginal expectation is the principal interest and correlation is a nuisance.

The results presented thus far are valid when there is a small and asymptotically diminishing fraction of missing data or when the data are missing completely at random in the sense of Rubin (1976). When the pattern of missing data at a given time depends on the previous outcomes, however, the consistency results for $\hat{\boldsymbol{\beta}}_R$ and for $\hat{\mathbf{V}}_R$ require that $\mathbf{R}$ be the true correlation matrix. Hence, the robustness to the choice of $\mathbf{R}$ does not hold in the case of nonrandom missing data. This is to be expected since most approaches will give inconsistent estimates when the data are not missing at random unless the assumed model is correct.

When there is little missing data or when the data are missing completely at random, choosing a correlation matrix, $\mathbf{R}_i$, involves a trade-off between the number of assumptions necessary to guarantee the consistency of $\hat{\boldsymbol{\beta}}$ and the efficiency of the estimate. If $\mathbf{R}_i = \mathbf{I}_{n_i}$, then $\hat{\boldsymbol{\beta}}_R$ and $\hat{\mathbf{V}}_R$ are consistent estimates of $\boldsymbol{\beta}$ and $\text{var}(\hat{\boldsymbol{\beta}}_R)$, respectively, as long as the choice of link function is correct. Subjects do not have to share the same correlation matrix. This is particularly desirable for binary outcomes where the range of permissible correlations depends on the propensity (Zeger et al., 1985). This choice, however, is less efficient than alternatives which explicitly account for correlation when the magnitude of association is large.

Several aspects of the GEE approach require further study. First, our results are asymptotic so it is important to assess the performance of GEEs for finite K. We are particularly interested in the performance of the robust variance estimate as it has application to a number of problems. Second, we have not considered systematic methods for choosing $\mathbf{R}$. While we prefer to explore several choices, residual analyses would be useful in selecting an optimal $\mathbf{R}$. Third, our approach can be extended to models that allow some severe imbalance due either to missing data or to different observation times. Quasi-likelihood analogues of random-effects models are of particular interest. Finally, as mentioned above, this approach can be extended to categorical data by developing reasonable working correlation matrices in this multivariate setting.

Acknowledgements

We are most grateful to Professor Cheryl Alexander, who collected and kindly made available the data discussed in Section 4.

Résumé

Les jeux de données longitudinales consistent en observations répétées d'un résultat et d'un ensemble de covariables pour chaque individu d'un ensemble. Un objectif de l'analyse statistique est de décrire l'espérance marginale de la variable résultat comme fonction des covariables en tenant compte de la corrélation entre les observations répétées concernant un même individu. Cet article propose une approche unifiée d'une telle analyse pour divers résultats discrets et continus.

Une classe d'équations généralisées d'estimation (EGE) pour les paramètres de regression est proposée. Les équations sont des extensions de celles utilisées dans les méthodes de quasi-vraisemblance (Wedderburn, 1974, *Biometrika* **61,** 439–447). Les EGE ont des solutions qui sont convergentes et asymptotiquement gaussiennes même quand la dépendance temporelle est spécifiée de façon erronée, comme on peut souvent s'y attendre. Un estimateur convergent de la variance est proposé. Nous illustrons l'utilisation de l'approche par les EGE avec des données longitudinales provenant d'une étude de l'effet du stress des mères sur la morbidité des enfants.

References

Anderson, D. A. and Aitkin, M. (1985). Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society, Series B* **47,** 203–210.

Azzalini, A. (1984). Estimation and hypothesis testing for collections of autoregressive time series. *Biometrika* **71,** 85–90.

Baker, R. J. and 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.

Fearn, T. (1975). A Bayesian approach to growth curves. *Biometrika* **62,** 89–100.

Grizzle, J. E. and Allen, D. M. (1969). Analysis of growth and dose response curves. *Biometrics* **25,** 357–381.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72,** 320–338.

Hui, S. L. (1984). Curve fitting for repeated measurements made at irregular time points. *Biometrics* **40,** 691–697.

Koch, G. C., Landis, J. R., Freeman, J. L., Freeman, D. H., and Lehmann, R. B. (1971). A general methodology for the analysis of repeated measurements of categorical data. *Biometrics* **33,** 133–158.

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

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38,** 963–974.

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

McCullagh, P. and Nelder, J. A. (1983). Quasi-likelihood functions. *Annals of Statistics* **11,** 59–67.

McCullagh, P. (1983). *Generalized Linear Models.* London: Chapman and Hall.

Ochi, Y. and Prentice, R. L. (1984). Likelihood inference in correlated probit regression. *Biometrika* **71,** 531–543.

Rao, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika* **52,** 447–458.

Rubin, D. B. (1976). Inference and missing data (with Discussion). *Biometrika* **63,** 581–592.

Stiratelli, R., Laird, N. and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics* **40,** 961–971.

Ware, J. H. (1985). Linear models for the analysis of several measurements in longitudinal studies. *American Statistician* **39,** 95–101.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61,** 439–447.

Zeger, S. L., Liang, K. Y., and Self, S. G. (1985). The analysis of binary longitudinal data with time-independent covariates. *Biometrika* **72,** 31–38.

Received April 1985; *revised November* 1985.