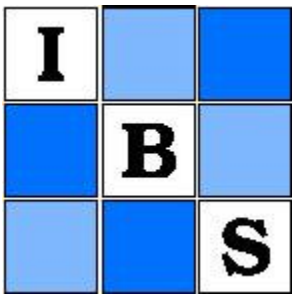

BIOMETRICS 46, 657–671
September 1990

Some Covariance Models for Longitudinal Count Data with Overdispersion

Peter F. Thall and Stephen C. Vail

Statistics/Computer & Information Systems Department, George Washington University, Washington, D.C. 20052, U.S.A.

SUMMARY

A family of covariance models for longitudinal counts with predictive covariates is presented. These models account for overdispersion, heteroscedasticity, and dependence among repeated observations. The approach is a quasi-likelihood regression similar to the formulation given by Liang and Zeger (1986, *Biometrika* **73,** 13–22). Generalized estimating equations for both the covariate parameters and the variance–covariance parameters are presented. Large-sample properties of the parameter estimates are derived. The proposed methods are illustrated by an analysis of epileptic seizure count data arising from a study of progabide as an adjuvant therapy for partial seizures.

1. Introduction

Consider a longitudinal data set consisting of a count response variable Y_{it} and a $p \times 1$ vector $\mathbf{X}_{it}$ of covariates observed at times $t = 1, \ldots, n_i$, for independent subjects $i = 1, \ldots, M$. Such data frequently arise in the clinical testing of new drugs, as well as other areas of application.

When there is a single response for each subject, i.e., all $n_i = 1$, generalized linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1983) are broadly applicable. However, data involving counts taken from biological units often exhibit variability exceeding that explained by exponential family probability models. In many circumstances, this can be modelled by unobserved random effects acting on the responses. The main problem faced by the scientist in such settings is identification of the linear predictor in the model. Several authors have stressed the importance of accounting for overdispersion and heteroscedasticity in order to correctly test elements of the linear component. The quasi-likelihood approach to this problem was first introduced by Wedderburn (1974).

For repeated outcomes, where $n_i > 1$, Liang and Zeger (1986) and Zeger and Liang (1986) have proposed quasi-likelihood models that describe the correlation structure among the responses, while also taking overdispersion into account. In the present article, we develop a means of postulating parametric forms for the within-subject covariance matrix and carrying out parameter estimation under the general framework of McCullagh (1983), and more specifically that of Liang and Zeger. We are motivated in part by circumstances wherein the nature and degree of the variability of the phenomenon over time may be as important as its average behavior.

For ease of notation we shall write $n_i = n$. Modifications to accommodate partially missing data are straightforward, provided they are missing completely at random. To account for regression of $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in})^{\mathrm{T}}$ on both time and the covariates, we denote

Key words: Covariance matrix; Generalized linear model; Longitudinal data; Overdispersion; Quasi-likelihood.

by $\mathbf{Z}_{it} = (Z_{it1}, \ldots, Z_{itp})^{\mathrm{T}}$ a vector consisting of $\mathbf{X}_{it}$ and possibly some functions of time, with link function g connecting each mean $\mu_{it} = \mathrm{E}(Y_{it})$ to the parameter vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^{\mathrm{T}}$ via $g(\mu_{it}) = \mathbf{Z}_{it}^{\mathrm{T}}\boldsymbol{\beta} = \eta_{it}$. In general, the visit times $t = 1, \ldots, n$ may be used to index unequally spaced study times $\tau_1 < \cdots < \tau_n$. This is useful for incorporating functions of time, e.g., $\beta_1 f_1(\tau) + \cdots + \beta_k f_k(\tau)$, within the linear component.

Denote $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in})^{\mathrm{T}}$, $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{in})^{\mathrm{T}}$, $\mathbf{S}_i = \mathbf{Y}_i - \boldsymbol{\mu}_i$, and $\mathbf{V}_i = \mathrm{cov}(\mathbf{Y}_i)$. We also require the covariate matrices $\mathbf{Z}_i = \partial\boldsymbol{\eta}_i/\partial\boldsymbol{\beta} = (\mathbf{Z}_{i1}, \ldots, \mathbf{Z}_{in})^{\mathrm{T}}$ and the derivative matrices $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta} = \boldsymbol{\Delta}_i\mathbf{Z}_i$, where $\boldsymbol{\Delta}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\eta}_i = \mathrm{diag}_t\{\partial\mu_{it}/\partial\eta_{it}\}$. The $\boldsymbol{\Delta}_i$'s are required to allow for links other than $\log(\mu_{it}) = \eta_{it}$, such as those obtained from the family suggested by Pregibon (1980). For convenience write $\mathbf{Y} = (\mathbf{Y}_1^{\mathrm{T}}, \ldots, \mathbf{Y}_M^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{S} = (\mathbf{S}_1^{\mathrm{T}}, \ldots, \mathbf{S}_M^{\mathrm{T}})^{\mathrm{T}}$, etc., and block-diagonal matrices $\mathbf{D} = \mathrm{diag}\{\mathbf{D}_i\}$ and $\mathbf{V} = \mathrm{diag}\{\mathbf{V}_i\}$, given specified $\mathbf{V}_i$. The p generalized estimating equations for $\boldsymbol{\beta}$ are

$$\sum_{i=1}^{M} \mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{S}_i = \mathbf{D}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{S} = \mathbf{0}. \tag{1}$$

Let $\boldsymbol{\alpha}$ be a $q \times 1$ vector of additional parameters arising in the formulation of $\mathbf{V}$, and denote $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}$.

The principal objective here is to present tractable parametric forms for $\mathbf{V} = \mathbf{V}(\boldsymbol{\theta})$ that account for heteroscedasticity, overdispersion relative to Poisson marginals, and dependence among the elements of each $\mathbf{Y}_i$. In addition to the assumed independence of the response vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_M$, we shall require that $\mathrm{E}(\mathbf{S}) = \mathbf{0}$, i.e., that the link g and linear components $\{\eta_{it}\}$ are correct. The joint distribution of the entries of each $\mathbf{Y}_i$ is specified only up to second moments. We derive $\mathbf{V}$ heuristically by including random effects for subject and time in each η_{it} and mixing over these effects. Estimation is carried out by alternating between solution of moment equations for the vector $\boldsymbol{\alpha}$ and the equations (1) for $\boldsymbol{\beta}$. An asymptotic distribution theory for the estimators then allows formal testing and subsequent remodelling of both $\boldsymbol{\mu}$ and $\mathbf{V}$, as functions of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Our formulation is analogous to that of Prentice (1988), who provides a quasi-likelihood framework for correlated binary responses with covariates. He derives the joint distribution of the regression and covariance parameters under a generalized estimating equation formulation that models both the pairwise correlations $\rho(\mathbf{Z})$ and the marginal probabilities $\pi(\mathbf{Z})$ as functions of the covariates.

In Section 2 we present a heuristic derivation of $\mathbf{V}$ based on random effects that act multiplicatively on the mean. Parameter estimation is described in Section 3, and a joint asymptotic distribution theory for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ is provided. In Section 4 we present several other forms of $\mathbf{V}$ suggested by the initial structure derived in Section 2. To illustrate our methods, an application to the analysis of epileptic seizure count data arising in a study of progabide as an adjuvant antiepileptic chemotherapy is presented in Section 5. Related methods are discussed in Section 6, including estimation of $\boldsymbol{\alpha}$ via pseudo-likelihood (Davidian and Carroll, 1987) and the quasi-likelihood approach of Prentice.

2. A Covariance Matrix

Denote $\sigma_{it}^2 = \mathrm{var}(Y_{it})$ $(1 \leq t \leq n)$ and $\sigma_{itu} = \mathrm{cov}(Y_{it}, Y_{iu})$ $(t < u)$. Consider the case of counts exhibiting extra-Poisson variation, in that $\sigma_{it}^2 > \mu_{it}$. This may be checked empirically by comparing the sample mean and variance of $\{Y_{it};\ 1 \leq i \leq M\}$ for each t.

We derive an initial form for $\mathbf{V}$ under the following assumptions. Let $\{\gamma_1, \ldots, \gamma_M\}$ be independent and identically distributed subject effects and $\{\zeta_1, \ldots, \zeta_n\}$ independent time effects, all unobserved and positive-valued. The two sets of effects are mutually independent. Conditional on these random effects, the responses are independent within subject and

each Y_{it} is Poisson with mean $\gamma_i \zeta_t \mu_{it}$ $(1 \leqslant i \leqslant M;\ 1 \leqslant t \leqslant n)$. In particular, $\mathrm{E}(Y_{it}) = \mu_{it}\mathrm{E}(\gamma_i \zeta_t)$. Since

$$\mathrm{var}(Y_{it}) = \mathrm{E}\{\mathrm{var}(Y_{it} \mid \gamma_i,\ \zeta_t)\} + \mathrm{var}\{\mathrm{E}(Y_{it} \mid \gamma_i,\ \zeta_t)\}$$

and

$$\mathrm{cov}(Y_{it},\ Y_{iu}) = \mathrm{E}\{\mathrm{E}(Y_{it} Y_{iu} \mid \gamma_i,\ \zeta_t,\ \zeta_u)\} - \mathrm{E}(Y_{it})\mathrm{E}(Y_{iu}),$$

it follows that

$$\begin{aligned} \sigma_{it}^2 &= \mu_{it}\mathrm{E}(\gamma_i \zeta_t) + \mu_{it}^2 \mathrm{var}(\gamma_i \zeta_t), \\ \sigma_{itu} &= \mu_{it}\mu_{iu}\mathrm{var}(\gamma_i)\mathrm{E}(\zeta_t \zeta_u). \end{aligned} \tag{2}$$

Assume the usual link $\log(\mu_{it}) = \eta_{it}$. Due to the mutual independence of all the random effects, $\mathrm{E}(\gamma_i \zeta_t) = \mathrm{E}(\gamma_t)\mathrm{E}(\zeta_t)$, $\mathrm{E}(\zeta_t \zeta_u) = \mathrm{E}(\zeta_t)\mathrm{E}(\zeta_u)$, and

$$\mathrm{var}(\gamma_i \zeta_t) = \mathrm{var}(\gamma_i)\mathrm{var}(\zeta_t) + \mathrm{var}(\gamma_i)\mathrm{E}^2(\zeta_t) + \mathrm{var}(\zeta_t)\mathrm{E}^2(\gamma_i).$$

For simplicity denote $\alpha_0 = \mathrm{var}(\gamma_1)/\mathrm{E}^2(\gamma_1)$ and $\delta_t = \mathrm{var}(\zeta_t)/\mathrm{E}^2(\zeta_t)$ $(1 \leqslant t \leqslant n)$. Incorporate the term $\log\{\mathrm{E}(\zeta_t)\}$ into the linear component η_{it} as an additive time effect and likewise absorb $\log\{\mathrm{E}(\gamma)\}$ into the constant term β_0. Under the reparameterization $\alpha_t = \delta_t(1 + \alpha_0)$, expression (2) now may be written as

$$\begin{aligned} \sigma_{it}^2 &= \mu_{it} + (\alpha_0 + \alpha_t)\mu_{it}^2, \\ \sigma_{itu} &= \alpha_0 \mu_{it}\mu_{iu}. \end{aligned} \tag{3}$$

Upon denoting $c_{it} = \mu_{it} + \alpha_t \mu_{it}^2$ and $\mathbf{C}_i = \mathrm{diag}_t\{c_{it}\}$, we can express the covariance matrix in the form

$$\mathbf{V}_i = \mathbf{C}_i + \alpha_0 \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}}, \tag{4}$$

which is clearly positive-definite. Further denoting $b_{it} = \mu_{it}^2/c_{it}$ and $b_{i\cdot} = b_{i1} + \cdots + b_{in}$, Householder's formula yields

$$\mathbf{V}_i^{-1} = \mathbf{C}_i^{-1} - \frac{\alpha_0}{1 + \alpha_0 b_{i\cdot}} \mathbf{C}_i^{-1} \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}} \mathbf{C}_i^{-1}, \tag{5}$$

which allows the estimating equations (1) to be expressed in a simple closed form.

There are three important simplifications and variants of this formulation. If $\alpha_1 = \cdots = \alpha_n = 0$, i.e., there are no time effects in $\mathbf{V}_i$, then (4) takes the form suggested by Nelder (1985, §5) for repeated measurements. When $\alpha_0 = 0$ there are no subject effects and we obtain the case of within-subject response independence with heteroscedastic overdispersion across time. Applying both constraints, $\alpha_1 = \cdots = \alpha_n$ and $\alpha_0 = 0$, yields the univariate additive overdispersion model of Breslow (1984). Covariance models similar to (4) are given by Morton (1987) in the context of nested strata, and Tsutakawa (1988) for analysis of lung cancer mortality rates with stratification by demographic groups and geographic region.

The heuristic derivation of $\mathbf{V}$ given above could be carried out equivalently by assuming a priori that all the random effects have mean 1, since $\mathrm{E}^2(\gamma_1)$ and $\mathrm{E}^2(\zeta_t)$ are absorbed into their respective variances. In this sense the formulation is analogous to that of the classical additive linear model for real-valued responses in which $\mathrm{E}(\gamma_1) = \mathrm{E}(\zeta_t) = 0$ and each Y_{it} has conditional mean $\mu_{it} + \gamma_i + \zeta_t$ and variance σ_ε^2, so that marginally

$$\mathbf{V}_i = \mathrm{diag}_t\{\mathrm{var}(\zeta_t) + \sigma_\varepsilon^2\} + \mathrm{var}(\gamma_1)\mathbf{J}, \tag{6}$$

where $\mathbf{J}$ denotes the $n \times n$ matrix of 1's. This corresponds to the usual linear model having

within- and between-subject variance components, but allowing heteroscedasticity over time. Thus, the two developments both can be carried out either from the full (Poisson or Gaussian) distributional assumption, or via quasi-likelihood with respective variance functions μ or 1. The variance matrix $\mathbf{V}_i$ in (6) does not vary with $\boldsymbol{\mu}_i$, however, while the analogous expression (4) derived under the Poisson assumption is an explicit function of $\boldsymbol{\mu}_i$.

3. Estimation

We approach the problem of parameter estimation by augmenting the p generalized estimating equations (1) in $\boldsymbol{\beta}$ with a second set of moment equations for the additional parameters $\boldsymbol{\alpha}$ arising in the formulation of $\mathbf{V}$. The underlying philosophy is to obtain a unified distribution theory for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, as in Moore (1986) and Prentice (1988), rather than to regard $\boldsymbol{\alpha}$ as a vector of nuisance parameters.

Denote $\mathbf{U}_{1i} = \mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{S}_i$ and define the q-vector of dispersion scores $\mathbf{U}_{2i} = (U_{2i1}, \ldots, U_{2iq})^{\mathrm{T}}$, where $q = n + 1$,

$$U_{2it} = \frac{S_{it}^2}{\sigma_{it}^2} - 1 \quad (1 \leqslant t \leqslant n),$$

and (7)

$$U_{2iq} = \sum_{1 \leqslant t < u \leqslant n} \left(\frac{S_{it}S_{iu}}{\mu_{it}\mu_{iu}} - \alpha_0 \right),$$

with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n, \alpha_0)^{\mathrm{T}}$. The $p + q$ estimating equations for $\boldsymbol{\theta}$ now may be expressed generally as

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix} = \sum_{i=1}^{M} \begin{bmatrix} \mathbf{U}_{1i} \\ \mathbf{U}_{2i} \end{bmatrix} = \mathbf{0}. \tag{8}$$

In practice, these equations are solved via a two-operation procedure that is iterated, as in Green (1984), Liang and Zeger (1986), Nelder and Pregibon (1987), and Prentice (1988). First $\mathbf{U}_1 = \mathbf{0}$ is solved for $\hat{\boldsymbol{\beta}}$ using a modified Fisher scoring approach; then $\hat{\boldsymbol{\alpha}}$ is obtained by solving the moment equations $\mathbf{U}_2 = \mathbf{0}$. The moment equations are solved directly when possible, otherwise via Newton–Raphson. These two operations are applied alternately until $\hat{\boldsymbol{\theta}}$ converges.

Initially $\mathbf{V}$ is set equal to the naive diagonal matrix of usual variances, assuming independence within and between subjects. That is, estimation of $\boldsymbol{\alpha}$ is suppressed. The estimates of $\boldsymbol{\beta}$ so obtained are those of a generalized linear model. Thereafter, $\mathbf{V}$ takes the form given by expression (4) and, using the previous value of $\hat{\boldsymbol{\beta}}$, the full parameter vector $\hat{\boldsymbol{\theta}}$ is computed iteratively until convergence. The convergence criterion used in the illustrative example given in Section 5 below is to stop at iteration $r + 1$ when $\max_j |(\hat{\theta}_j^{(r+1)} - \hat{\theta}_j^{(r)})/s(\hat{\theta}_j^{(r)})| < 10^{-4}$.

The statistic $H = \det\{\mathrm{var}(\hat{\boldsymbol{\theta}})^{-1}\}$ is employed here to assess goodness of fit, where $\det(\cdot)$ denotes the determinant function. As noted by Godambe and Heyde (1987), H is an increasing function of the squared vector correlation $\rho^2 = \det\{\mathrm{E}(\mathbf{U}\mathbf{U}_*^{\mathrm{T}})\}/\{\det(\boldsymbol{\Lambda})\det(\boldsymbol{\Lambda}_*)\}$ between $\mathbf{U}$ and the score vector $\mathbf{U}_*$ of the "true" likelihood, denoting $\boldsymbol{\Lambda}_* = \mathrm{cov}(\mathbf{U}_*)$. We use H because $\mathbf{U}_*$ is of course unknown, hence ρ^2 cannot be computed. The numerical values of $\log(H)$ that are given in Table 4 of the illustrative example thus may be interpreted only ordinally, in that a larger value corresponds to a higher value of ρ^2, hence a better fit to the data. Note that a generalized Pearson statistic based on $\sum \mathbf{S}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{S}_i$ or $\sum (S_{it}^2/\sigma_{it}^2)$ would be rather misleading here due to the essential role of the moment equations in estimating $\mathbf{V}$.

Define the $(p+q)\times(p+q)$ matrices $\mathbf{G} = M^{-1}(-\partial\mathbf{U}/\partial\boldsymbol{\theta})$, $\boldsymbol{\Gamma} = \mathrm{E}(\mathbf{G})$, and $\boldsymbol{\Lambda} = M^{-1}\mathrm{cov}(\mathbf{U})$, and partition $\mathbf{G}$ as

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{bmatrix} = \frac{1}{M}\begin{bmatrix} -\partial\mathbf{U}_1/\partial\boldsymbol{\beta} & -\partial\mathbf{U}_1/\partial\boldsymbol{\alpha} \\ -\partial\mathbf{U}_2/\partial\boldsymbol{\beta} & -\partial\mathbf{U}_2/\partial\boldsymbol{\alpha} \end{bmatrix},$$

with $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$ partitioned in like manner. Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$. The following theorems are proved in the Appendix, under suitable regularity conditions.

Theorem 1 There exists a sequence $\{\hat{\boldsymbol{\theta}}_M\}$ such that (a) $\Pr\{\mathbf{U}(\hat{\boldsymbol{\theta}}_M) = \mathbf{0}\} \to 1$ and (b) $\hat{\boldsymbol{\theta}}_M \to \boldsymbol{\theta}_0$ in probability. This sequence is unique in that for any $\{\tilde{\boldsymbol{\theta}}_M\}$ satisfying (a) and (b), it must be the case that $\Pr\{\tilde{\boldsymbol{\theta}}_M = \hat{\boldsymbol{\theta}}_M\} \to 1$.

Theorem 2 $M^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix consistently estimated by $[\boldsymbol{\Gamma}(\hat{\boldsymbol{\theta}})]^{-1}\boldsymbol{\Lambda}(\hat{\boldsymbol{\theta}})\{[\boldsymbol{\Gamma}(\hat{\boldsymbol{\theta}})]^{-1}\}^{\mathrm{T}}$.

This result provides a joint asymptotic distribution for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$. In this sense it may be regarded as an extension of Liang and Zeger (1986, Theorem 2), with the important difference that we assume $\mathbf{V}$ is correct while they do not. From another viewpoint, our results are a multivariate version of those given by Moore (1986). Prentice (1988) provides a similar distribution theory for the case of binary responses, but estimates the second moment parameters in a different manner.

For computation of the covariance matrix of $\hat{\boldsymbol{\theta}}$, first write $\boldsymbol{\Lambda}_{11} = M^{-1}\sum \boldsymbol{\Lambda}_{11,i}$, etc. Since $\boldsymbol{\Lambda}_{11,i} = \mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{D}_i$, it follows from (5) that

$$\Lambda_{11,i,(r,s)} = \sum_{t=1}^{n} Z_{itr}Z_{its}\frac{\Delta_{it}^2}{c_{it}} - \frac{\alpha_0}{1+\alpha_0 b_{i\cdot}}\sum_{t=1}^{n} Z_{itr}\frac{\Delta_{it}\mu_{it}}{c_{it}}\sum_{u=1}^{n} Z_{ius}\frac{\Delta_{iu}\mu_{iu}}{c_{iu}}, \tag{9}$$

$(1 \leqslant r, s \leqslant p)$. The submatrices $\boldsymbol{\Lambda}_{12,i} = \mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathrm{E}(\mathbf{S}_i\mathbf{U}_{2i}^{\mathrm{T}}) = \boldsymbol{\Lambda}_{21,i}^{\mathrm{T}}$ and $\boldsymbol{\Lambda}_{22,i} = \mathrm{E}(\mathbf{U}_{2i}\mathbf{U}_{2i}^{\mathrm{T}})$ involve mixed central moments of $\{Y_{it},\ 1 \leqslant t \leqslant n\}$, up to fourth order. In general, these matrices may be estimated by substituting $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ and $Y_{it} - \hat{\mu}_{it}$ for S_{it} throughout. Substantial simplifications may be made using the first- and second-order moment assumptions $\mathrm{E}(S_{it}) = 0$ and $\mathrm{E}(\mathbf{S}_i\mathbf{S}_i^{\mathrm{T}}) = \mathbf{V}_i$. For $\boldsymbol{\Gamma}$, it is easily verified that $\boldsymbol{\Gamma}_{11} = M^{-1}\sum \mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{D}_i$ and $\boldsymbol{\Gamma}_{12} = \mathbf{0}$. The remaining components of $\boldsymbol{\Gamma}_i$ are

$$\boldsymbol{\Gamma}_{21,i} = \begin{bmatrix} \mathrm{diag}_t\left[\dfrac{\partial\{\log(\sigma_{it}^2)\}}{\partial\eta_{it}}\right] \\ (n-1)\alpha_0\left(\dfrac{\Delta_{i1}}{\mu_{i1}}, \ldots, \dfrac{\Delta_{in}}{\mu_{in}}\right) \end{bmatrix}\mathbf{Z}_i \tag{10}$$

and

$$\boldsymbol{\Gamma}_{22,i} = \begin{bmatrix} & & \mu_{i1}^2/\sigma_{i1}^2 \\ \mathrm{diag}_t\{\mu_{it}^2/\sigma_{it}^2\} & & \vdots \\ & & \mu_{in}^2/\sigma_{in}^2 \\ 0 & \cdots\ 0 & n(n-1)/2 \end{bmatrix}, \tag{11}$$

using formula (4), with $\hat{\boldsymbol{\theta}}$ in place of $\boldsymbol{\theta}$.

An immediate consequence of Theorem 2 is that $M^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is asymptotically normal with covariance matrix $\boldsymbol{\Lambda}_{11}(\boldsymbol{\theta})^{-1}$, since $\boldsymbol{\Lambda}_{11} = \boldsymbol{\Gamma}_{11}$. Liang and Zeger (1986) prove asymptotic normality for $\hat{\boldsymbol{\beta}}$ under the more general assumptions that $\mathbf{V}$ may be incorrect but

$M^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) = O_p(1)$. While this condition is difficult to verify when the model for second-order structure used in $\mathbf{U}_2$ and $\mathbf{V}$ is not assumed to be correct, the definition of $\boldsymbol{\Lambda}$ shows that in any case the covariance matrix of $\hat{\boldsymbol{\beta}}$ may be correctly estimated by

$$\boldsymbol{\Gamma}_{11}^{-1}\left[\frac{1}{M}\sum_{i=1}^{M}\mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{S}_i\mathbf{S}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{D}_i\right]\boldsymbol{\Gamma}_{11}^{-1}, \tag{12}$$

using $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ throughout. As noted by Prentice (1988) in the analogous formulation for binary responses, the choice between (12) and $\boldsymbol{\Lambda}_{11}^{-1}$ is essentially a trade-off between robustness and precision, provided that in the latter case one has done a good job of modelling $\mathbf{V}$.

4. Other Forms of V

In application, the modelling process may require several forms of $\mathbf{V}$, each motivating its own version of $\mathbf{U}_2$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\Gamma}$. Under the special case of the general formulation given by (3), (4) where $\alpha_0 = 0$ and $q = n$, corresponding to independence within subjects, the $(n+1)$st entry of $\mathbf{U}_2$ is deleted. Here $\mathbf{V}$ is diagonal, with $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$ simplified in fairly obvious ways. In this case, the model for the marginal distribution at each occasion is equivalent to the formulation given by Breslow (1984).

For the case $\alpha_1 = \cdots = \alpha_n$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_0)^{\mathrm{T}}$, the first n entries of $\mathbf{U}_{2i}$ collapse to the single term $\sum_t \{(S_{it}^2/\sigma_{it}^2) - 1\}$, while $\sigma_{it}^2 = \mu_{it} + (\alpha_1 + \alpha_0)\mu_{it}^2$. Here $q = 2$,

$$\boldsymbol{\Gamma}_{21,i} = \begin{bmatrix} \dfrac{\partial\{\log(\sigma_{i1}^2)\}}{\partial\eta_{i1}}, & \cdots, & \dfrac{\partial\{\log(\sigma_{in}^2)\}}{\partial\eta_{in}} \\ (n-1)\alpha_0\left(\dfrac{\Delta_{i1}}{\mu_{i1}}, \ldots, \dfrac{\Delta_{in}}{\mu_{in}}\right) & & \end{bmatrix}\mathbf{Z}_i,$$

$$\boldsymbol{\Gamma}_{22,i} = \begin{bmatrix} \sum_{t=1}^{n}\dfrac{\mu_{it}^2}{\sigma_{it}^2} & \sum_{t=1}^{n}\dfrac{\mu_{it}^2}{\sigma_{it}^2} \\ 0 & \dfrac{n(n-1)}{2} \end{bmatrix},$$

and the remaining submatrices of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$ are obtained essentially as before.

The general expression (4) for $\mathbf{V}_i$ suggests the alternative $q = (n+1)$-parameter model in which

$$\sigma_{it}^2 = \alpha_t\mu_{it} + \alpha_0\mu_{it}^2 \quad \text{and} \quad \sigma_{itu} = \alpha_0\mu_{it}\mu_{iu}, \tag{13}$$

so that overdispersion is represented both additively and multiplicatively. This form of the variance is suggested in unpublished work by Nelder for the univariate case ($n = 1$), where he notes that it was used by Bartlett as early as 1936 to model insect count data. Here $\mathbf{V}_i = \mathbf{C}_i + \alpha_0\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}$ as before, but now $\mathbf{C}_i = \mathrm{diag}_t\{\alpha_t\mu_{it}\}$. If $\alpha_1 = \cdots = \alpha_n$ this reduces to the $q = 2$-parameter formulation where $\boldsymbol{\alpha} = (\alpha_1, \alpha_0)^{\mathrm{T}}$, $\sigma_{it}^2 = \alpha_1\mu_{it} + \alpha_0\mu_{it}^2$, and the matrices $\mathbf{U}_{2i}$, $\boldsymbol{\Gamma}_{2i}$, and $\boldsymbol{\Lambda}_i$ are of the same general form as given previously for the corresponding $q = 2$-parameter version of (4). The further constraint $\alpha_1 = 1$ yields the model suggested by Nelder (1985) in the context of grouped count data.

The requirement that all $\sigma_{itu} > 0$ may be relaxed by setting

$$\sigma_{itu} = \rho_{tu}\sigma_{it}\sigma_{iu}$$

for some suitable form of σ_{it}^2 and correlation ρ_{tu}. This is essentially the correlation

formulation proposed by Liang and Zeger (1986, §3), but with time-varying overdispersion. Zeger (1988) also presents a regression model for a time series of counts with similar correlation structure. This formulation may be more desirable in settings where correlations that vary with (t, u) are appropriate, since the assumption that the ζ_t's are independent does not provide such structure. We discuss this point further in Section 6. Although formally we must constrain $q \leqslant n(n+1)/2$, in the present context a more parsimonious approach, e.g., all $\rho_{tu} = \rho$ or $\rho_{|t-u|}$, is highly desirable. In the former case, U_{2iq} is as in expression (7) but with $\sigma_{it}\sigma_{iu}$ in place of $\mu_{it}\mu_{iu}$ and α_0 now the common correlation parameter. Writing $\mathbf{A}_i = \text{diag}_t\{\sigma_{it}^2\}$, for these models we have

$$\mathbf{V}_i = (1 - \alpha_0)\mathbf{A}_i + \alpha_0 \mathbf{A}_i^{1/2}\mathbf{J}\mathbf{A}_i^{1/2}$$

and

$$\mathbf{V}_i^{-1} = (1 - \alpha_0)^{-1}\left\{\mathbf{A}_i^{-1} - \frac{\alpha_0}{1 + \alpha_0(n-1)}\,\mathbf{A}_i^{-1/2}\mathbf{J}\mathbf{A}_i^{-1/2}\right\}. \tag{14}$$

For example, given the additive overdispersion variance $\sigma_{it}^2 = \mu_{it} + \alpha_t\mu_{it}^2$, both $\mathbf{U}_1$ and $\mathbf{\Gamma}_{11}$ are computed using (14). The other submatrices of $\mathbf{\Lambda}$ are also obtained empirically as before, $\mathbf{\Gamma}_{21,i}$ is of the form (10) but with the first matrix having $(n+1, t)$ entry $\{(n-1)\alpha_0/2\}[\partial\{\log(\sigma_{it}^2)\}/\partial\eta_{it}]$ rather than $(n-1)\alpha_0\Delta_{it}/\mu_{it}$ $(t = 1, \ldots, n)$, and

$$\mathbf{\Gamma}_{22,i} = \begin{bmatrix} & & 0 \\ & \text{diag}_t\left\{\dfrac{\mu_{it}^2}{\sigma_{it}^2}\right\} & \vdots \\ & & 0 \\ \dfrac{\alpha_0(n-1)}{2}\left(\dfrac{\mu_{i1}^2}{\sigma_{i1}^2}, \ldots, \dfrac{\mu_{in}^2}{\sigma_{in}^2}\right) & & n(n-1)/2 \end{bmatrix}.$$

The various forms of $\mathbf{V}$ described above are presented systematically in Table 1. The first two entries in column 2 of the table are the most general models (13) and (4), respectively, with columns 1 and 3 containing their respective independence and correlation versions. In the latter case, α_0 denotes the correlation. For convenience we shall refer to each model by its two-digit row–column location in the table, e.g., model 21 has $\sigma_t^2 = \mu_t + \alpha_t\mu_t^2$ and $\sigma_{tu} = 0$. The third and fourth rows are obtained by setting $\alpha_1 = \cdots = \alpha_n$ in each of the first

Table 1
Formulations for the within-subject covariance matrix $\mathbf{V}$*, suppressing the subject index* i *for simplicity. Each cell contains the variance* σ_t^2 *and covariance* σ_{tu} *of the model.*

	1	2	3
1	$\alpha_t\mu_t$	$\alpha_t\mu_t + \alpha_0\mu_t^2$	$\alpha_t\mu_t$
	0	$\alpha_0\mu_t\mu_u$	$\alpha_0\sigma_t\sigma_u$
2	$\mu_t + \alpha_t\mu_t^2$	$\mu_t + (\alpha_t + \alpha_0)\mu_t^2$	$\mu_t + \alpha_t\mu_t^2$
	0	$\alpha_0\mu_t\mu_u$	$\alpha_0\sigma_t\sigma_u$
3	$\alpha_1\mu_t$	$\alpha_1\mu_t + \alpha_0\mu_t^2$	$\alpha_1\mu_t$
	0	$\alpha_0\mu_t\mu_u$	$\alpha_0\sigma_t\sigma_u$
4	$\mu_t + \alpha_1\mu_t^2$	$\mu_t + (\alpha_1 + \alpha_0)\mu_t^2$	$\mu_t + \alpha_1\mu_t^2$
	0	$\alpha_0\mu_t\mu_u$	$\alpha_0\sigma_t\sigma_u$
5	μ_t	$\mu_t + \alpha_0\mu_t^2$	μ_t
	0	$\alpha_0\mu_t\mu_u$	$\alpha_0\sigma_t\sigma_u$

Table 2

Successive two-week seizure counts for 59 *epileptics. Covariates are adjuvant treatment* (0 = *placebo*, 1 = *progabide*), *eight-week baseline seizure counts, and age* (*in years*).

ID	Y_1	Y_2	Y_3	Y_4	Trt	Base	Age
104	5	3	3	3	0	11	31
106	3	5	3	3	0	11	30
107	2	4	0	5	0	6	25
114	4	4	1	4	0	8	36
116	7	18	9	21	0	66	22
118	5	2	8	7	0	27	29
123	6	4	0	2	0	12	31
126	40	20	23	12	0	52	42
130	5	6	6	5	0	23	37
135	14	13	6	0	0	10	28
141	26	12	6	22	0	52	36
145	12	6	8	4	0	33	24
201	4	4	6	2	0	18	23
202	7	9	12	14	0	42	36
205	16	24	10	9	0	87	26
206	11	0	0	5	0	50	26
210	0	0	3	3	0	18	28
213	37	29	28	29	0	111	31
215	3	5	2	5	0	18	32
217	3	0	6	7	0	20	21
219	3	4	3	4	0	12	29
220	3	4	3	4	0	9	21
222	2	3	3	5	0	17	32
226	8	12	2	8	0	28	25
227	18	24	76	25	0	55	30
230	2	1	2	1	0	9	40
234	3	1	4	2	0	10	19
238	13	15	13	12	0	47	22
101	11	14	9	8	1	76	18
102	8	7	9	4	1	38	32
103	0	4	3	0	1	19	20
108	3	6	1	3	1	10	30
110	2	6	7	4	1	19	18
111	4	3	1	3	1	24	24
112	22	17	19	16	1	31	30
113	5	4	7	4	1	14	35
117	2	4	0	4	1	11	27
121	3	7	7	7	1	67	20
122	4	18	2	5	1	41	22
124	2	1	1	0	1	7	28
128	0	2	4	0	1	22	23
129	5	4	0	3	1	13	40
137	11	14	25	15	1	46	33
139	10	5	3	8	1	36	21
143	19	7	6	7	1	38	35
147	1	1	2	3	1	7	25
203	6	10	8	8	1	36	26
204	2	1	0	0	1	11	25
207	102	65	72	63	1	151	22
208	4	3	2	4	1	22	32
209	8	6	5	7	1	41	25
211	1	3	1	5	1	32	35
214	18	11	28	13	1	56	21
218	6	3	4	0	1	24	41
221	3	5	4	3	1	16	32
225	1	23	19	8	1	22	26
228	2	3	0	1	1	25	21
232	0	0	0	0	1	13	36
236	1	4	3	2	1	12	37

two rows, respectively. In particular, 31 is the usual multiplicative overdispersion model while 41 is the additive overdispersion model treated by Williams (1982), Breslow (1984), Moore (1986), and Lawless (1987). The fifth row is obtained from row 1 by setting $\alpha_t = 1$ $(1 \leqslant t \leqslant n)$, so that 52 is Nelder's (1985) proposal, while 51 is the usual generalized linear model.

In solving the moment equations for the central model 22, we use $(M - p)/M$ in place of 1 in U_{2it} $(1 \leqslant t \leqslant n)$, and $[\{Mn(n-1)/2\} - p]/\{Mn(n-1)/2\}$ as a multiplier of α_0 in U_{2iq}, to correct for estimation of $\boldsymbol{\beta}$, with these modifications also included in $\boldsymbol{\Lambda}$. Similar modifications apply for the other models. For numerical values of $\hat{\alpha}_t$ that are negative, the estimator is defined to be either 0 or 1 as α_t is included additively or multiplicatively in σ_{it}^2.

5. Example

As an illustration, we present analyses of data arising from a clinical trial of 59 epileptics carried out by Leppik et al. (1985). The data are given in Table 2.

Patients suffering from simple or complex partial seizures were randomized to receive either the antiepileptic drug progabide or a placebo, as an adjuvant to standard chemotherapy. At each of four successive postrandomization clinic visits, the number of seizures occurring over the previous 2 weeks was reported. Although each patient subsequently was crossed over to the other treatment, we shall consider only the four precrossover responses. As shown in Table 3, the seizure counts exhibit a high degree of extra-Poisson variation, heteroscedasticity, and within-patient dependence.

Each of the first six covariance models given in Table 1 was fit to the data, with the additive overdispersion model 41 also included for comparison. The covariates appearing in the fitted models are baseline seizure rate, computed as the logarithm of $\frac{1}{4}$ the 8-week prerandomization seizure count, logarithm of age in years, and the binary indicators Trt for the progabide group and Visit_4 for the fourth clinic visit. These were obtained from a larger set of predictors via preliminary stepdown procedures carried out for each model.

Table 4 gives the parameter estimates, their standard errors, and the value of $\log(H)$ for each model. Convergence was obtained in between five and eight iterations of the two-stage estimation procedure for all models shown. For these data, the original model 22 that was derived heuristically in Section 2 gives the best fit, although model 12 and the correlation model 23 fit the data almost as well. The common properties of these three

Table 3

Summary statistics for the two-week seizure counts. Within each group, the first column contains the mean and variance at each visit, followed by the correlations.

	Placebo ($M_1 = 28$)					Progabide ($M_2 = 31$)				
Visit	$\bar{Y}$ s^2	Correlations				$\bar{Y}$ s^2	Correlations			
1	9.36 102.76	1.00				8.58 332.72	1.00			
2	8.29 66.66	.78	1.00			8.42 140.65	.91	1.00		
3	8.79 215.29	.51	.66	1.00		8.13 193.05	.91	.92	1.00	
4	7.96 58.18	.67	.78	.68	1.00	6.71 126.88	.97	.95	.95	1.00

models are that, in addition to their accounting for the rather strong within-subject covariance, the time-varying overdispersion is parameterized additively in their variances.

The interaction between treatment and baseline seizure rate in these analyses produces a rather interesting result—namely, the predicted mean seizure rate for the progabide group is either higher or lower than that for the placebo group, accordingly as the baseline count does or does not exceed a critical threshold. This threshold varies between 11.2 and 15.2 seizures per 2-week period for the seven models considered, with the highest value corresponding to model 22. This suggests that progabide may be contraindicated for patients with high seizure rates. We regard this as a qualitative result, however, since it is based on a single data set and a particular family of models.

A visual scan of the data suggests that progabide patient #207 is an outlier, since both the baseline and posttreatment counts are much higher than those of the other patients. Deletion of this patient produces a marked drop in the means, variances, and correlations within the progabide group (Table 5), although the extra-Poisson variation and within-subject dependence are still substantial. While a comparison of these adjusted summary statistics with those of the placebo group suggests a greater treatment effect, deletion of this patient from the data has no clinical basis.

Plots of the standardized residuals $e_{it} = (Y_{it} - \hat{\mu}_{it})/\hat{\sigma}_{it}$ for model 22 on the log baseline counts at each occasion (Figure 1) are perhaps more telling. In addition to the fact that each of the plots shows a random scatter, none of the residuals $\mathbf{e} = (e_1, e_2, e_3, e_4) = (2.06, .58, .70, 1.32)$ of patient #207 are extreme. The largest residual at $t = 1$ is that of placebo patient #135, for whom $\mathbf{e} = (3.84, 2.99, .75, -1.32)$. This is due in part to the marked improvement of this patient over the course of the trial, as evidenced by the monotone

Table 4

Covariate and dispersion parameter estimates for each of seven covariance matrix models. Standard deviations of the estimates are given in parentheses.

Model	11	12	13	21	22	23	41
Int	−2.695	−1.456	−2.753	−1.590	−1.350	−1.735	−1.456
	(.902)	(.933)	(.907)	(.907)	(.904)	(.915)	(.927)
Base	.933	.870	.922	.892	.877	.888	.896
	(.087)	(.105)	(.085)	(.107)	(.103)	(.098)	(.123)
Trt	−1.439	−.987	−1.532	−.983	−.958	−1.063	−.907
	(.418)	(.423)	(.419)	(.412)	(.390)	(.418)	(.412)
Base.Trt	.595	.372	.629	.375	.352	.401	.351
	(.171)	(.209)	(.172)	(.203)	(.196)	(.206)	(.204)
Age	.895	.567	.923	.589	.531	.639	.542
	(.264)	(.272)	(.267)	(.264)	(.266)	(.268)	(.266)
Visit_4	−.168	−.170	−.170	−.156	−.159	−.155	−.150
	(.065)	(.070)	(.066)	(.077)	(.072)	(.077)	(.080)
$\hat{\alpha}_1$	3.641	1.936	3.575	.372	.148	.364	.452
	(.924)	(.126)	(.887)	(.151)	(.138)	(.143)	(.139)
$\hat{\alpha}_2$	4.768	2.769	4.846	.638	.383	.664	
	(1.558)	(.196)	(1.629)	(.276)	(.249)	(.296)	
$\hat{\alpha}_3$	8.250	5.749	8.255	.777	.546	.786	
	(3.656)	(.423)	(3.682)	(.310)	(.298)	(.319)	
$\hat{\alpha}_4$	2.456	1.000	2.479	.218	.000	.224	
	(.443)	(.079)	(.447)	(.080)	(.078)	(.081)	
$\hat{\alpha}_0$		.225	.348		.227	.334	
		(.056)	(.070)		(.055)	(.069)	
$\log(H)$	26.98	47.55	32.29	41.46	47.96	46.75	29.11

decline in seizure count. The largest residual of 4.48 at $t = 2$ is due to progabide patient #225, while the two extreme residuals at $t = 3$ of 3.20 and 4.17 correspond to patients #112 and #227, respectively. In each instance, a relatively large value of e_{it} is due to some deviation within subject from the predicted pattern over the four occasions. This phenomenon will arise almost invariably in any setting where a predictive model is fit to longitudinal data. The model thus accounts for the high dispersion in these data quite well, since none

Table 5
Summary statistics for the two-week seizure counts of the progabide group, after deletion of patient #207

Visit	$\bar{Y}$ / s^2	Correlations			
1	5.47				
	33.22	1.00			
2	6.53				
	31.43	.45	1.00		
3	6.00				
	54.34	.63	.70	1.00	
4	4.83				
	18.35	.77	.72	.83	1.00

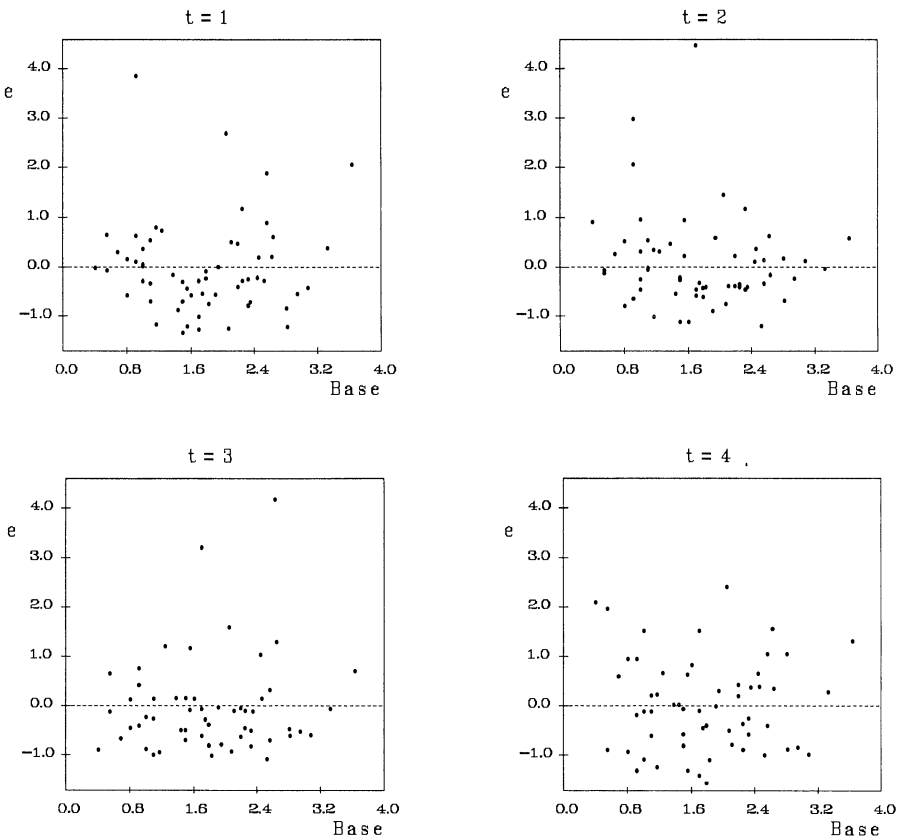

Figure 1. Plots of the residuals $e_{it} = (Y_{it} - \hat{\mu}_{it})/\hat{\sigma}_{it}$ on log-baseline seizure count, by occasion.

of the extreme residuals was due to a value of the response variable being large relative to the sample.

6. Discussion

Quasi-likelihood regression models for longitudinal data recently have received considerable attention in the literature. As an alternative to the generalized quasi-likelihood approach where $\mathbf{V}$ is modelled parametrically, Wei and Stram (1988) propose modelling only the marginal distributions while estimating the within-subject covariance empirically. Other treatments are given by McCullagh (1983), Jorgensen (1983, 1987), Green (1984), and Godambe and Heyde (1987).

In the univariate response setting ($n = 1$), several approaches that allow modelling of both $\mathrm{E}(Y)$ and $\mathrm{var}(Y)$ have been proposed. Nelder and Pregibon (1987) present a generalized quasi-likelihood that provides a general framework for parametric modelling of the variance. In addition to the usual link $g(u) = \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta} = \eta$, their formulation allows $\mathrm{var}(Y) = \phi\mathbf{V}_{\theta}(\mu)$ with θ linked to ϕ via $h(\phi) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\theta}$, and an algorithm for estimating $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is described. Davidian and Carroll (1987) treat the heterogeneous regression model where $\mathrm{E}(Y) = f(\mathbf{Z}, \boldsymbol{\beta})$ and $\mathrm{var}(Y) = \sigma^2 g^2(\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\alpha})$ for given covariates $\mathbf{Z}$, $\mathbf{X}$ and unknown parameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and $\sigma^2 > 0$, but no distributional assumptions beyond the forms of the first and second moments are made. They also focus attention on the variance function, proposing that $\boldsymbol{\alpha}$ be estimated by maximizing the so-called "pseudo-likelihood" obtained by conditioning on $\hat{\boldsymbol{\beta}}$ and assuming a Gaussian form.

A referee has suggested estimating $\boldsymbol{\alpha}$ in the present setting by adopting a pseudo-likelihood approach using the multivariate normal, and also via quasi-likelihood. For the first method, letting $\mathrm{var}(\mathbf{Y}_i) = \sigma^2\mathbf{V}_i$ for some covariance matrix $\mathbf{V}_i$, the pseudo-log-likelihood is

$$l(\boldsymbol{\alpha}, \sigma) = -M \log(\sigma) - \frac{1}{2}\sum_{i=1}^{M} [\log\{\det(\mathbf{V}_i)\} + \sigma^{-2}\mathbf{S}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{S}_i]. \tag{15}$$

Although minimization of (15) is apparently more involved than the naive moment estimation scheme given in Section 3, pseudo-likelihood may prove especially tractable for patterned covariance matrices having inverse and determinant that are computable in closed form. Again suppressing the index i for simplicity, each matrix corresponding to a cell in Table 1 is of the general form $\mathbf{V} = \mathbf{C} + \alpha\mathbf{a}\mathbf{a}^{\mathrm{T}}$ with $\mathbf{C} = \mathrm{diag}\{c_t\}$, so that

$$\log\{\det(\mathbf{V})\} = \log(1 + \alpha \sum a_t^2 c_t^{-1}) + \sum \log(c_t)$$

and

$$\mathbf{S}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{S} = \mathbf{S}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{S} - \frac{\alpha\mathbf{S}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{a}\mathbf{a}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{S}}{1 + \alpha\mathbf{a}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{a}}$$

$$= \sum(S_t^2 c_t^{-1}) - \frac{\alpha(\sum a_t S_t c_t^{-1})^2}{1 + \alpha \sum a_t^2 c_t^{-1}}.$$

Thus, differentiation of (15) is straightforward. For application of this approach to estimation of $\boldsymbol{\alpha}$ in the generalized estimating equation regime of Liang and Zeger (1986), where $\mathbf{V} = \mathbf{A}^{1/2}\mathbf{R}(\boldsymbol{\alpha})\mathbf{A}^{1/2}$ with $\mathbf{A}$ the diagonal matrix of usual generalized linear model variances, the same sort of consideration would apply—namely, the determinant and inverse of the "working correlation matrix" $\mathbf{R}(\boldsymbol{\alpha})$ should be computable in closed form.

Prentice (1988) estimates covariance matrix parameters in the case of correlated binary responses via a second set of generalized estimating equations, solved along with those for $\boldsymbol{\beta}$. This suggests yet a third approach for estimation of $\boldsymbol{\alpha}$ in the present setting. To apply Prentice's method here, let $\mathbf{S}_i^*$ denote the vector of dimension $n(n + 1)/2$ composed

of the entries

$$\frac{S_{it}S_{iu}}{(\mu_{it}\mu_{iu})^{1/2}} - \delta_{itu},$$

$1 \leqslant t \leqslant u \leqslant n$, where $\delta_{itu} = \sigma_{itu}/(\mu_{it}\mu_{iu})^{1/2}$. For $\mathbf{E}_i = d\boldsymbol{\delta}_i/d\boldsymbol{\alpha}$ and $\mathbf{W}_i$ the identity or a suitable working correlation matrix, the generalized estimating equations for $\boldsymbol{\alpha}$ are

$$\sum_{i=1}^{M} \mathbf{E}_i^{\mathrm{T}}\mathbf{W}_i^{-1}\mathbf{S}_i^* = \mathbf{0}.$$

These alternative methods for estimation of $\boldsymbol{\alpha}$ raise a number of important issues, including those pertaining to comparisons of the statistical properties of the estimators under various formulations of $\mathbf{V}$. Although such general questions of how second-order parameters are to be estimated are quite important, we do not pursue this issue further here.

As noted earlier, the correlation version of each model in column 2 of Table 1 allows one to account for the times (t, u) in modelling $\mathrm{cov}(\mathbf{Y}_t, \mathbf{Y}_u)$, in addition to allowing these terms to be negative. This also may be accomplished by specifying a nondiagonal parametric form for $\mathrm{cov}(\boldsymbol{\zeta})$ in the original formulation given in Section 2, rather than assuming that $\zeta_1, \ldots, \zeta_n$ are independent. If we denote $\mathbf{V}_0 = \mathrm{diag}_t(\mu_t)$, $\alpha_0 = \mathrm{var}(\gamma_1)$, and assume for simplicity that all the random effects have mean 1, then reasoning as before, we obtain the generalization of (4) given by

$$\mathbf{V} = \mathbf{V}_0 + \mathbf{V}_0[(\alpha_0 + 1)\mathrm{cov}(\boldsymbol{\zeta}) + \alpha_0\mathbf{J}]\mathbf{V}_0.$$

This provides a broader framework for modelling $\mathbf{V}$, although we advocate as parsimonious a parameterization of $\mathrm{cov}(\boldsymbol{\zeta})$ as is reasonable in a given setting.

The primary aim of the present article has been to propose a family of parametric models for the covariance matrix in the quasi-likelihood regression framework for longitudinal count data. As the above generalization shows, more complex versions of the models used here to fit the epilepsy data certainly are available, and quite possibly may give better fits. The data analyses presented in Section 5 are intended to provide a reasonably simple illustration of the general methodology, however, so that we have intentionally avoided using more elaborate versions of $\mathbf{V}$. Still, the models presented in Table 1 comprise a broad collection of within-subject covariance structures, while each parameterization of $\mathbf{V}$ is rather parsimonious. Given the distributional results for both $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, this provides a practical means for modelling both $\boldsymbol{\mu}$ and $\mathbf{V}$ in the generalized estimating equation setting.

Acknowledgements

The authors thank Jia-Yeong Tsay of the U.S. National Institutes of Health for providing the data set used in the example, and N. Younes for assisting with the residual plots. We also thank the referees for their valuable comments and suggestions. This work was partially supported by the National Institute of Diabetes, Digestive and Kidney Diseases Grant #R01-AM-35952.

Résumé

On présente une famille de modèles de covariance pour des comptages répétés, avec des covariables prédictives. Ces modèles analysent la dispersion totale, l'hétéroscédasticité, et la dépendance entre mesures répétées. L'approche est semblable à la régression par quasi-vraisemblance formulée par Liang et Zeger (1986, *Biometrika* **73,** 13–22). Une généralisation des équations est faite pour l'estimation des coefficients des covariables et des variances–covariances. On en déduit un grand nombre de propriétés des estimateurs. Les méthodes proposées sont appliquées à des données de comptages de crises d'épilepsie survenues pendant des essais de régulateurs utilisés au cours d'une thérapie de complément dans des crises légères.

References

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics* **33,** 38–44.
Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association* **82,** 1079–1091.
Foutz, R. V. (1977). On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association* **72,** 147–148.
Godambe, V. P. and Heyde, C. C. (1987). Quasi-likelihood and optimal estimation. *International Statistical Review* **55,** 231–244.
Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society, Series B* **46,** 149–192.
Jorgensen, B. (1983). Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika* **70,** 19–28.
Jorgensen, B. (1987). Exponential dispersion models (with Discussion). *Journal of the Royal Statistical Society* **49,** 127–162.
Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics* **15,** 209–225.
Leppik, I. E., et al. (1985). A double-blind crossover evaluation of progabide in partial seizures. *Neurology* **35,** 285.
Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.
McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics* **11,** 59–67.
McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models.* London: Chapman and Hall.
Moore, D. F. (1986). Asymptotic properties of moment estimators for overdispersed counts and proportions. *Biometrika* **73,** 583–588.
Morton, R. (1987). A generalized linear model with nested strata of extra-Poisson variation. *Biometrika* **74,** 247–257.
Nelder, J. A. (1985). Quasi-likelihood and GLIM. In *Generalized Linear Models*, R. Gilchrist, B. Francis, and J. Whittaker (eds), 120–127. Berlin: Springer-Verlag.
Nelder, J. A. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika* **74,** 221–232.
Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135,** 370–384.
Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied Statistics* **29,** 15–24.
Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44,** 1033–1048.
Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.
Tsutakawa, R. K. (1988). Mixed model for analyzing geographic variability in mortality rates. *Journal of the American Statistical Association* **83,** 37–42.
Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61,** 439–447.
Wei, L. J. and Stram, D. (1988). Analyzing repeated measurements with possibly missing observations by modelling marginal distributions. *Statistics in Medicine* **7,** 139–148.
Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics* **31,** 144–148.
Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42,** 121–130.
Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika* **75,** 621–629.

Received January 1989; *revised July and September* 1989; *accepted September* 1989.

Appendix

Our strategy for deriving the distribution of $\hat{\theta}$ follows Moore (1986), who treats the case of univariate counts and proportions with additive overdispersion, where $\text{var}(Y_i) = \nu_1(\mu_i) + \alpha\nu_2(\mu_i)$. His approach is suitably modified here to accommodate the multivariate quasi-likelihood framework and multidimensional α. First, uniform consistency of $\mathbf{G}(\theta)$ for $\boldsymbol{\Gamma}(\theta)$ in a neighborhood of the true value θ_0 is established, and this is applied together with the inverse function theorem to obtain consistency of $\hat{\theta}$

for $\boldsymbol{\theta}$, as in Foutz (1977). These results are then used in conjunction with the asymptotic normality of $M^{-1/2}\mathbf{U}(\boldsymbol{\theta}_0)$ to establish asymptotic normality for $M^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. We assume that

(i) $M^{-1}\sum_{i=1}^{M}\boldsymbol{\Lambda}_i \to \boldsymbol{\Lambda}^0$,
(ii) $\sup_{i,t,r}|Z_{itr}| < \infty$,
(iii) $\boldsymbol{\beta}$ lies in a compact subset of $\mathscr{E}^p$,
(iv) $\sup \mathrm{E}|S_{it_1}\cdot\cdots\cdot S_{it_6}| < \infty$, i.e., all sixth-order moments are finite, and
(v) $\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)$ is positive-definite,

where the supremum in (iv) is over all i and $1 \leq t_1, \ldots, t_6 \leq n$, allowing repetition. All limits are taken as $M \to \infty$.

Lemma 1 Under conditions (i)–(iv), $M^{-1/2}\mathbf{U}(\boldsymbol{\theta}_0)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Lambda}^0$.

Lemma 2 Under condition (ii), for any suitably small open neighborhood $\mathscr{N}_0$ of $\boldsymbol{\theta}_0$,

$$\sup_{\boldsymbol{\theta}\in\mathscr{N}_0}\|\mathbf{G}(\boldsymbol{\theta}) - \boldsymbol{\Gamma}(\boldsymbol{\theta})\| \to 0$$

in probability, where $\|\cdot\|$ denotes the spectral norm.

Proof of Lemma 1 Under condition (i), by Theorem B of Serfling (1980, p. 30), it suffices to show that

$$M^{-1}\sum_{i=1}^{M}\mathrm{E}\{|\mathbf{U}_i|^2 I(|\mathbf{U}_i| > \varepsilon M^{1/2})\} \to 0, \tag{A.1}$$

where $|\cdot|$ denotes Euclidean norm and $I(\cdot)$ is the indicator function. Applying Minkowski's inequality, condition (A.1) is implied by

$$\sup_{i,r}\mathrm{E}|U_{1ir}|^3 < \infty, \quad \sup_{i,t}\mathrm{E}|U_{2it}|^3 < \infty. \tag{A.2}$$

Write

$$U_{1ir} = \sum_{t=1}^{n}\left[Z_{itr}\Delta_{it} - \mu_{it}\frac{\alpha_0}{1+\alpha_0 b_{i\cdot}}\left(\sum_{u=1}^{n}\frac{Z_{iur}\Delta_{iu}\mu_{iu}}{c_{iu}}\right)\right]\frac{S_{it}}{c_{it}} \tag{A.3}$$

($1 \leq r \leq p$), and recall the specific form of $\mathbf{U}_{2i}$. Since (ii) and (iii) imply $\{\eta_{it}\}$ is uniformly bounded, these conditions together with (iv) imply (A.2).

Proof of Lemma 2 It suffices to show componentwise convergence. Denoting the coefficient of S_{it} in (A.3) by ω_{itr}, we may write

$$(\mathbf{G} - \Gamma)_{11,i,(r,s)} = -\sum_{t=1}^{n}\left(\frac{\partial\omega_{itr}}{\partial\beta_s}\right)S_{it}, \quad (1 \leq r, s \leq p).$$

Since $\boldsymbol{\theta}$ lies in a compact set, under condition (ii) direct evaluation shows $\{\partial\omega_{itr}/\partial\beta_s\}$ is bounded uniformly in i, t, r, and s. The same assumptions imply $\{\eta_{it}\}$ and hence $\{\sigma_{it}^2\}$ are uniformly bounded, so it follows by the strong law that

$$M^{-1}\sum_{i=1}^{M}\left(\frac{\partial\omega_{itr}}{\partial\beta_s}\right)S_{it} \to 0$$

with probability 1 for each t, r, and s. Similar reasoning may be applied to the entries of the other three submatrices of $\mathbf{G} - \boldsymbol{\Gamma}$.

Theorem 1 is proved in a manner similar to the proof given by Foutz (1977, Theorem 2). The key requirements are the positive-definiteness of $\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)$ in the limit and the uniform convergence of $\mathbf{G}$ to $\boldsymbol{\Gamma}$ on $\mathscr{N}_0$.

Theorem 2 is obtained via a straightforward Taylor expansion of $\mathbf{U}(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$. Utilizing the uniform boundedness of $\{\partial^2 U_{ik}/\partial\theta_l\partial\theta_l\}$, the consistency of $\hat{\boldsymbol{\theta}}$ and of $\mathbf{G}$, and the positive-definiteness of $\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)$, the asymptotic distribution of $M^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is obtained from that of $M^{-1/2}\mathbf{U}(\boldsymbol{\theta}_0)$.