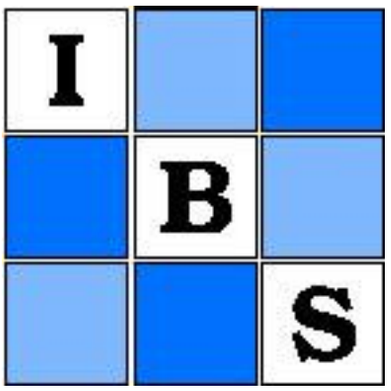

BIOMETRICS 47, 825–839
September 1991

Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses

Ross L. Prentice* and Lue Ping Zhao

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, Washington 98104, U.S.A.

SUMMARY

Generalized estimating equations are introduced in an ad hoc fashion for the covariance matrix of a multivariate response. These equations are to be solved jointly with score equations from a generalized linear model for mean parameters. A class of quadratic exponential models is used to develop joint estimating equations for mean and covariance parameters in a more systematic fashion, and proposals for the use of such equations are developed. Comments on the relative merits of the ad hoc and model-based approaches to estimation are given and a regression illustration with a bivariate response is provided.

1. Introduction

Many biometrical applications involve multivariate response vectors with interest often focusing on the dependence of the mean response on associated covariates or experimental conditions. The structure of the covariance matrix and the dependence of covariances on covariates may also be of substantive interest, for example, in family studies or longitudinal studies. Analyses based on multivariate normal models have been proposed for such problems, largely because means and covariances are readily estimated. There has been considerable work (e.g., Whittle, 1961; Crowder, 1985) exploring the properties of normal-theory-based estimation when the response variables may be decidedly nonnormal, with results indicating that means and covariances are generally consistently estimated, but the estimating efficiency may be poor and conventional standard error estimates will be incorrect.

Much effort has been directed to addressing the estimation of parameters in the mean, for a univariate response. This work has focussed on estimation under the generalized linear model, and on quasilikelihood extensions that use the score equations from generalized linear models as estimating equations for mean parameters, along with an ad hoc estimator of an overdispersion parameter (e.g., McCullagh and Nelder, 1989). This approach was extended to the estimation of parameters in a multivariate mean by Liang and Zeger (1986) and Zeger and Liang (1986). These authors also suggested that one may be able to improve the efficiency of the estimators of mean parameters by simultaneously estimating parameters in the covariance matrices for the response vector. For the special case of binary data, Prentice (1988) formalized this idea somewhat by introducing a second set of estimating equations for parameters in the correlation matrix. This idea is extended here to mean and covariance parameter estimation for a general multivariate response including discrete and continuous responses and mixtures thereof.

---

*To whom requests for reprints should be addressed.

Key words: Correlation; Covariance; Estimating equations; Quadratic exponential model; Regression; Score equations.

In order to gain insight into the properties of estimators solving certain estimating equations, and possibly to identify estimators with better properties, the topic of mean and covariance parameter estimation is then taken up in the context of a quadratic exponential family of distributions. Such a family has been shown in the econometrics literature (Gourieroux, Montfort, and Trognon, 1984) to be unique in yielding consistent estimates of mean and covariance parameters under mild regularity conditions, even if sampling takes place from outside the family. This approach generates a class of estimating equations that are readily applied and that are completely analogous to Liang and Zeger's results for mean parameters.

Before going into these derivations, some discussion of our emphasis on marginal moments seems appropriate. Models for means and covariances can typically be sensibly specified, even if the observed response vectors include a variable number of elements. Mean and covariance parameters can often be defined to address the questions of primary data-analytic interest, and such parameters are usually readily interpreted. In Section 3 we will discuss a class of models for a mixed discrete and continuous response vector which, like the multivariate normal model, is parametrized in terms of the mean vector and covariance matrix of the response.

This class of models contrasts with many other multivariate models, including log-linear models for discrete response (e.g., Bishop, Fienberg, and Holland, 1975), which are expressed in terms of parameters whose interpretation derives from the conditional distribution of each element of the response vector given the values of the remaining elements. While such parameters will be of interest in some applications, their modelling and interpretation are likely to suffer severely if the dimension of the response vector is variable. Along the same lines, such "conditional" models are generally not reproductive, so that a different modelling assumption is made concerning the *marginal* distribution of a subset of a response vector, according to whether other elements of the response are simultaneously modelled. In comparison, random-effects models in which elements of the response vector are regarded as statistically independent given one or more hypothetical "latent" variables (e.g., Manski and McFadden, 1981) will generally possess a desirable reproductive property. However, key parameters usually must be interpreted conditional on the latent variables, distributional assumptions on the latent variables are necessarily somewhat arbitrary and untestable, and computational aspects of model fitting may be difficult.

An illustration may help to distinguish the modelling assumptions mentioned above. A study of an educational intervention program to prevent cigarette smoking among elementary school children in Washington State is one of several group randomized trials being coordinated by the Fred Hutchinson Cancer Research Center. Forty school districts have been randomized to either intervention or control, and students are followed from third through tenth grade in order to ascertain cigarette smoking habits by means of self-report and biochemical analysis. Because of the group randomization it is critical that study evaluation accommodate possible correlation among the responses among students in the same school district. For example, the response for a student could be a binary variate taking value unity if the student was smoking regularly by the end of some fixed time from randomization and value zero otherwise, or could be a graded response defined according to the frequency and duration of cigarette smoking. Of primary interest is the effect of intervention activities on smoking probability, or probabilities. Intervention effects can be conveniently modelled as regression effects on the response mean when using the same models and analyses that would be employed if all responses were independent. In these analyses correlations among the smoking habits of students in the same school district play a nuisance role. In addition, a regression model for the pairwise correlations, or covariances, of the response could be used to examine whether intervention activities altered the dispersion among the smoking habits of students in the same school district. In contrast, intervention effects defined in terms of models for the smoking habits of individual students, given the corresponding smoking habits of all students in the same school district, would seem to be difficult indeed

to interpret. If an intervention effect could not be identified one could not readily distinguish an ineffective intervention from an intervention whose effect has been diminished or removed by conditioning on the smoking habits of the other students. Such conditional models may, however, have a role in attempting to explain, rather than identify, an intervention effect. A latent variable approach to this problem would allow each school district to have a distinct parameter—for example, a location parameter in a binary regression model for smoking probability—and would assume that such parameters are independent random variables from some distribution. Regression coefficients that assess intervention effectiveness then are to be interpreted relative to these hypothetical random effects, in contrast to those based on regression models for the mean response, which have an unconditional, population-averaged interpretation (e.g., Zeger, Liang, and Albert, 1988). A given random-effects model and distributional assumption implies a corresponding covariance structure on the response vector. This structure may involve a parametrization that is not sufficiently flexible or interpretable, especially if the covariances are of substantive interest.

These types of considerations suggest that models that emulate the multivariate normal distribution in that they are parametrized in terms of the response mean and covariance, in conjunction with suitable estimation procedures, could play a major role in multivariate data analysis.

2. Ad Hoc Estimating Equations for Means and Covariances

Consider a sample of K independent random observations $\mathbf{y}_k^t = (y_{k1}, \ldots, y_{kn_k})$, $k = 1, \ldots, K$, of a general multivariate response vector. As a multivariate extension of quasilikelihood methods, Liang and Zeger (1986) and Zeger and Liang (1986) propose that a parameter vector $\boldsymbol{\beta}$ in the mean response $\boldsymbol{\mu}_k^t = \boldsymbol{\mu}_k^t(\boldsymbol{\beta}) = \{\mathrm{E}(y_{k1}), \mathrm{E}(y_{k2}), \ldots\}$ be estimated as solution to

$$K^{-1/2} \sum_{k=1}^{K} D_{k11}^t V_{k11}^{-1} (\mathbf{y}_k - \boldsymbol{\mu}_k) = \mathbf{0}, \tag{1}$$

where $D_{k11} = \partial \boldsymbol{\mu}_k / \partial \boldsymbol{\beta}^t$ and V_{k11} is the variance matrix for $\mathbf{y}_k$. They suggest further that parameters $\boldsymbol{\alpha}$ that characterize $V_{k11} = V_{k11}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ can be estimated using simple functions of residuals. Prentice (1988) extended this idea in the context of binary response vectors by introducing a second set of estimating equations for $\boldsymbol{\alpha}$. These equations were of the same form as (1) but with $\mathbf{y}_k - \boldsymbol{\mu}_k$ replaced by the vector of differences between the empirical and true pairwise correlations. Asymptotic normality results were provided for the joint estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

This simple estimating equation approach for means and covariances applies equally to response variables other than binary. For convenience of notation let

$$\boldsymbol{\sigma}_k^t = (\sigma_{k11}, \sigma_{k12}, \ldots, \sigma_{k22}, \ldots, \sigma_{kn_k n_k})$$

denote the covariance matrix V_{k11} in vector form. Let

$$\mathbf{s}_k^t = (s_{k11}, s_{k12}, \ldots, s_{kn_k n_k})$$

denote the corresponding empirical covariance vector defined by

$$s_{kij} = s_{kij}(\boldsymbol{\beta}) = (y_{ki} - \mu_{ki})(y_{kj} - \mu_{kj}).$$

For a general response vector, which may for example include both discrete and continuous elements, one can then estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ as solutions to (1) and

$$K^{-1/2} \sum_{k=1}^{K} D_{k22}^{t} V_{k22}^{-1} (\mathbf{s}_k - \boldsymbol{\sigma}_k) = \mathbf{0}, \tag{2}$$

where $D_{k22}^{t} = \partial \boldsymbol{\sigma}_k / \partial \boldsymbol{\alpha}^{t}$ and V_{k22} is a "working" variance matrix for the vector of empirical covariances $\mathbf{s}_k$.

In order to specify the asymptotic distribution for $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}$ that solve (1) and (2), denote

$$D_k = \begin{pmatrix} \partial \boldsymbol{\mu}_k / \partial \boldsymbol{\beta}^{t} & 0 \\ 0 & \partial \boldsymbol{\sigma}_k / \partial \boldsymbol{\alpha}^{t} \end{pmatrix} = \begin{pmatrix} D_{k11} & D_{k12} \\ D_{k21} & D_{k22} \end{pmatrix},$$

$$V_k = \begin{pmatrix} V_{k11} & 0 \\ 0 & V_{k22} \end{pmatrix}, \quad \mathbf{f}_k = \begin{pmatrix} \mathbf{y}_k - \boldsymbol{\mu}_k \\ \mathbf{s}_k - \boldsymbol{\sigma}_k \end{pmatrix},$$

and $\boldsymbol{\Sigma} = K^{-1} \sum_{k=1}^{K} D_k^{t} V_k^{-1} D_k$. As shown in Appendix 1, arguments analogous to those given in Liang and Zeger (1986) and Prentice (1988) then show that $\{K^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{t}, K^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})^{t}\}$ quite generally has an asymptotic normal distribution, as $K \to \infty$, with mean zero and with covariance matrix consistently estimated by

$$K^{-1} \boldsymbol{\Sigma}^{-1} \left(\sum_{k=1}^{K} D_k^{t} V_k^{-1} \mathbf{f}_k \mathbf{f}_k^{t} V_k^{-1} D_k \right) \boldsymbol{\Sigma}^{-1}. \tag{3}$$

Furthermore, $K^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ has an asymptotic normal distribution with mean zero and variance consistently estimated by

$$K \left(\sum_{k=1}^{K} D_{k11}^{t} V_{k11}^{-1} D_{k11} \right)^{-1} \left(\sum_{k=1}^{K} D_{k11}^{t} V_{k11}^{-1} (\mathbf{y}_k - \boldsymbol{\mu}_k)(\mathbf{y}_k - \boldsymbol{\mu}_k)^{t} V_{k11}^{-1} D_{k11} \right)$$
$$\times \left(\sum_{k=1}^{K} D_{k11}^{t} V_{k11}^{-1} D_{k11} \right)^{-1}, \tag{4}$$

regardless of whether $V_{k11} = \operatorname{var} \mathbf{y}_k$, $k = 1, \ldots, K$, has been correctly specified. In (3) and (4) all quantities are evaluated at $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$. Also the asymptotic distribution for $\hat{\boldsymbol{\beta}}$ does not depend on the precision with which the covariance parameter $\boldsymbol{\alpha}$ is estimated, provided only that the estimator of $\boldsymbol{\alpha}$ is $K^{1/2}$-consistent. Note that an alternative model-based estimator of the variance of $K^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is obtained by replacing $(\mathbf{y}_k - \boldsymbol{\mu}_k)(\mathbf{y}_k - \boldsymbol{\mu}_k)^{t}$ in (4) by V_{k11}.

These features suggest that simple special cases of (1) and (2) will be attractive if interest resides primarily in the mean parameter $\boldsymbol{\beta}$. One can specify, or build, a model for the variance matrix $V_{k11}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ that may imply good efficiency for $\hat{\boldsymbol{\beta}}$ estimation without being overly concerned about the choice of weight matrices V_{k22} for $\mathbf{s}_k$, $k = 1, \ldots, K$. For example, it may often be good enough, as far as $\boldsymbol{\beta}$ estimation is concerned, to take V_{k22} to be diagonal, giving rise to a very simple computational approach.

On the other hand, if substantive interest resides in both the mean and covariance parameters, a simple specification of (2) may yield parameter estimates that are unacceptably inefficient for $\boldsymbol{\alpha}$. Hence a systematic means of generating estimating equations for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ would be desirable. Even for $\boldsymbol{\beta}$ estimation such a systematic approach may allow for more effective specification of the variance matrix for the response vector, and hence more efficient inference on mean parameters, especially if the sample is not large.

3. Estimating Equations Arising from a Quadratic Exponential Model

Estimating equations for parameters in the mean and covariance vectors of the response $\mathbf{y}_k$ can be generated under a quadratic exponential model

$$\Pr_k(\mathbf{y}_k; \mu_k, \sigma_k) = \Delta_k^{-1} \exp\{\mathbf{y}_k^t \theta_k + \mathbf{w}_k^t \lambda_k + c_k(\mathbf{y}_k)\}, \tag{5}$$

where $\mathbf{w}_k^t = (y_{k1}^2, y_{k1}y_{k2}, \ldots, y_{k2}^2, y_{k2}y_{k3}, \ldots)$, $c_k(\cdot)$ is a "shape" function, $\Delta_k = \Delta_k\{\theta_k, \lambda_k, c_k(\cdot)\}$ is a normalization constant, and where the "canonical" parameters $\theta_k^t = \theta_k^t(\mu_k, \sigma_k) = (\theta_{k1}, \ldots, \theta_{kn_k})$ and $\lambda_k^t = \lambda_k^t(\mu_k, \sigma_k) = (\lambda_{k11}, \lambda_{k12}, \ldots, \lambda_{k22}, \lambda_{k23}, \ldots)$ are expressed as functions of the "marginal" parameters (μ_k, σ_k). Gourieroux et al. (1984) consider (5), and under mild regularity show maximum likelihood estimation under any specified shape function, $c_k(\cdot)$, to yield consistent estimates of mean and covariance parameters. Moreover, they show this class of models to be unique in implying such consistency. The Jacobian of the transformation from (θ_k, λ_k) to (μ_k, σ_k) is simply the inverse of the covariance matrix for $(\mathbf{y}_k^t, \mathbf{w}_k^t)$, so that this transformation will generally be one-to-one unless the distribution of $(\mathbf{y}_k^t, \mathbf{w}_k^t)$ is degenerate.

The class of models (5) seems important for multivariate data analysis. Any member of this class, determined by specifying $c_k(\cdot)$, provides a fully parametric model for a multivariate discrete, continuous, or mixed response that involves only the response mean and covariance. The special case given by $c_k(\cdot) \equiv 0$ reduces to the multivariate normal distribution if all response variables are continuous on the entire real line. This special case also gives a conditional normal distribution for continuous response variables given corresponding discrete variables. As such it is closely related to the interesting graphical models of Lauritzen and Wermuth (1989) and Frydenberg and Lauritzen (1989). However, these authors evidently specify parameters for the marginal distribution of the discrete response and the conditional distribution of the continuous given the discrete, whereas (5) is parametrized in terms of marginal means and covariances for the entire response vector. If the shape function $c_k(\cdot)$ is regarded as a parameter then the entire class of models (5) is easily shown to be reproductive. Certain special cases, such as the multivariate normal distribution, are also reproductive. The class of models (5) appears to be quite rich for any given type of response vector, provided the shape function is allowed to vary.

Zhao and Prentice (1990) have considered score estimating equations under a model of the form (5) for multivariate binary data. For a more general response vector, after some algebra given in Appendix 2, the score equations for β and α, under any special case of (5), can be written

$$K^{-1/2} \sum_{k=1}^{K} D_k^t V_k^{-1} \mathbf{f}_k = \mathbf{0}, \tag{6}$$

where

$$D_k = \begin{pmatrix} \partial \mu_k / \partial \beta^t & 0 \\ \partial \sigma_k / \partial \beta^t & \partial \sigma_k / \partial \alpha^t \end{pmatrix},$$

$$V_k = \begin{pmatrix} V_{k11} & V_{k12} \\ V_{k21} & V_{k22} \end{pmatrix} = \begin{pmatrix} \operatorname{var} \mathbf{y}_k & \operatorname{cov}(\mathbf{y}_k, \mathbf{s}_k) \\ \operatorname{cov}(\mathbf{s}_k, \mathbf{y}_k) & \operatorname{var} \mathbf{s}_k \end{pmatrix}, \tag{7}$$

and $\mathbf{f}_k$ is as above. The corresponding information matrix is given by

$$K^{-1} \sum_{k=1}^{K} D_k^t V_k^{-1} D_k. \tag{8}$$

Note that efficient estimation of parameters in the mean vector and covariance matrix under the class of models (5) depends on the shape function $c_k(\cdot)$, and on the type of response vector, only through the third and fourth central moments of $\mathbf{y}_k$, $k = 1, \ldots, K$. This is analogous to efficient estimation of the mean parameters under a generalized linear model involving only the variance of the sampling distribution.

A pseudo-maximum likelihood approach (Gourieroux et al., 1984) to the estimation of mean and covariance parameters would proceed by specifying shape functions $c_k(\cdot)$, $k = 1, \ldots, K$, and solving (6). With few exceptions such solution requires a double iterative procedure. Trial values of $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$ specify $\boldsymbol{\mu}_k(\boldsymbol{\beta}_0)$ and $\boldsymbol{\sigma}_k(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)$, $k = 1, \ldots, K$. An iterative calculation is then required to compute corresponding canonical parameters $\boldsymbol{\theta}_{k0}$ and $\boldsymbol{\lambda}_{k0}$ for each $k = 1, \ldots, K$. These values then allow the third and fourth moments in V_k to be directly calculated by summation or integration thereby giving updated values $\boldsymbol{\beta}_u$ and $\boldsymbol{\alpha}_u$ according to

$$\begin{pmatrix} \boldsymbol{\beta}_u \\ \boldsymbol{\alpha}_u \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\alpha}_0 \end{pmatrix} + \left(\sum D_{k0}^t V_{k0}^{-1} D_{k0}\right)^{-1} \left(\sum D_{k0}^t V_{k0}^{-1} \mathbf{f}_{k0}\right), \tag{9}$$

where all quantities on the right side are evaluated at $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)$. This double iterative procedure is typically computationally unattractive. For example, if each element of $\mathbf{y}_k$ can take m possible values, the calculation of its third and fourth moments involves a summation over m^{n_k} terms. Numerical integration may be required for these moment calculations if some components of $\mathbf{y}_k$ are continuous.

A more convenient estimation procedure arises by specifying "working" variance matrices in (6) in which the third and fourth moments in V_k are expressed as functions of $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k$, $k = 1, \ldots, K$. This approach allows $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ to be calculated using the simple iterative procedure (9). Conceptually these estimators are equally attractive to pseudo-maximum likelihood estimates, in view of the correspondence between $c_k(\cdot)$ and the third and fourth moments in V_k, $k = 1, \ldots, K$. Upon noting that working specifications of these moments could include additional parameters for which $K^{1/2}$-consistent estimates are inserted, one has a procedure for estimating mean and covariance parameters that naturally generalizes Liang and Zeger's (1986) results for mean parameters to mean and covariance parameters.

Under either the pseudo-maximum likelihood or the estimating equation approach, the asymptotic distribution of $\{K^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t, K^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})^t\}$ quite generally is normal with mean zero, and with variance consistently estimated by (3) (Appendix 1). If the V_k matrices used to generate scores obtain, then (3) can be replaced by a presumably more stable, model-based variance estimator given by the inverse of (8).

Consider now some possible working variance matrices V_k, $k = 1, \ldots, K$. A working matrix V_k will include the variance matrix var $\mathbf{y}_k = V_{k11}$, equivalent to $\boldsymbol{\sigma}_k(\boldsymbol{\beta}, \boldsymbol{\alpha})$, as the upper diagonal block, along with off-diagonal block ($V_{k12} = V_{k21}^t$) and lower diagonal block specifications, for cov$(\mathbf{y}_k, \mathbf{s}_k)$ and var $\mathbf{s}_k$, respectively. The following specifications may be considered:

(i) *Independence working matrices* Assuming the elements of $\mathbf{y}_k$ to be independent gives cov$(\mathbf{y}_k, \mathbf{s}_k) \equiv 0$ and a diagonal matrix var $\mathbf{s}_k$ with entries var $s_{kij} = \sigma_{kii}\sigma_{kjj}$ for $i \neq j$. The specification can be completed, somewhat arbitrarily, by inserting the normal theory value var $s_{kii} = 2\sigma_{kii}^2$.

The facts that var $\mathbf{s}_k$, which is of dimension $n_k(n_k + 1)/2$, is diagonal and V_k is block-diagonal allows a particularly simple computational procedure (9) even if certain n_k's are fairly large. This working specification may be quite adequate for a rather broad range of applications in which the dependencies among elements of $\mathbf{y}_k$ are not strong, and may be adequate for mean parameter estimation even more broadly.

(ii) *Gaussian working matrices* A normal distribution for $\mathbf{y}_k$ gives, for all i, j, l, m,

$$\text{cov}(y_{ki}, s_{kjl}) = \text{E}\{(y_{ki} - \mu_{ki})(y_{kj} - \mu_{kj})(y_{kl} - \mu_{kl})\} = 0$$

and

$$\begin{aligned}\operatorname{cov}(s_{kij}, s_{klm}) &= \mathrm{E}\{(y_{ji} - \mu_{ki})(y_{kj} - \mu_{kj})(y_{kl} - \mu_{kl})(y_{km} - \mu_{km})\} - \sigma_{kij}\sigma_{klm} \\ &= \sigma_{kil}\sigma_{kjm} + \sigma_{kim}\sigma_{kjl}.\end{aligned}$$

This specification can be expected to yield estimates of high efficiency even if dependencies are strong, provided the joint sampling distribution does not differ substantially from normality. Expression (3) replaces the standard normal-theory-based variance estimators by estimators that adapt to departures from normality.

(iii) *Gaussian matrices with common third and fourth-order correlations* The Gaussian variance matrices can be generalized by setting

$$\mathrm{E}\{(y_{ki} - \mu_{ki})(y_{kj} - \mu_{kj})(y_{kl} - \mu_{kl})\} = \gamma_{ijl}(\sigma_{kii}\sigma_{kjj}\sigma_{kll})^{1/2},$$

and

$$\begin{aligned}&\mathrm{E}\{(y_{ki} - \mu_{ki})(y_{kj} - \mu_{kj})(y_{kl} - \mu_{kl})(y_{km} - \mu_{km})\} \\ &\quad = \sigma_{kij}\sigma_{klm} + \sigma_{kil}\sigma_{kjm} + \sigma_{kim}\sigma_{kjl} + \delta_{ijlm}(\sigma_{kii}\sigma_{kjj}\sigma_{kll}\sigma_{kmm})^{1/2},\end{aligned}$$

where γ_{ijl} and δ_{ijlm}, $i \leqslant j \leqslant l \leqslant m$, are additional parameters to be estimated. Natural $K^{1/2}$-consistent estimators $\hat{\gamma}_{ijl}$ and $\hat{\delta}_{ijlm}$ are given, respectively, by the average over all k such that $n_k \geqslant l$ of

$$(y_{ki} - \mu_{ki})(y_{kj} - \mu_{kj})(y_{kl} - \mu_{kl})(\sigma_{kii}\sigma_{kjj}\sigma_{kll})^{-1/2},$$

and by the average over all k such that $n_k \geqslant m$ of

$$\begin{aligned}&\{(y_{ki} - \mu_{ki})(y_{kj} - \mu_{kj})(y_{kl} - \mu_{kl})(y_{km} - \mu_{km}) - (\sigma_{kij}\sigma_{klm} + \sigma_{kil}\sigma_{klm} + \sigma_{kim}\sigma_{kjl})\} \\ &\quad \times (\sigma_{kii}\sigma_{kjj}\sigma_{kll}\sigma_{kmm})^{-1/2},\end{aligned}$$

with all quantities evaluated at $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$.

This generalization allows the estimation procedure to adapt to skewness or kurtosis in the sampling distribution, relative to the normal distribution. This procedure may have quite good asymptotic efficiency properties, but sample sizes (K) may need to be large for such asymptotic properties to be realized. Simpler specifications in which certain of the γ_{ijl} and δ_{ijlm} parameters are restricted to be equal may often be more useful, particularly if the elements of $\mathbf{y}_k$ are in some sense exchangeable.

Various other working specifications of V_k, $k = 1, \ldots, K$, could be entertained; for example, higher-order correlations could also be added to the specification (i). In each case the estimators $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ will generally be consistent and asymptotically normal with robust variance estimator (3), and with efficiency properties that can be expected to be good if the third and fourth central moments of the response have been fairly well specified and estimated.

The above discussion presumes the mean and covariance models $\boldsymbol{\mu}_k(\boldsymbol{\beta})$, $\boldsymbol{\sigma}_k(\boldsymbol{\beta}, \boldsymbol{\alpha})$, $k = 1, \ldots, K$, to have been correctly specified. The ad hoc estimating equations of Section 2, however, yield consistent estimators of $\boldsymbol{\beta}$ even if the covariance models have been misspecified. In order that $\boldsymbol{\beta}$ estimation be robust to covariance misspecification, one should refrain from using the postulated covariance model form to strengthen the estimation of $\boldsymbol{\beta}$. Hence it is natural to replace $D_{k21} = \partial\boldsymbol{\sigma}_k(\boldsymbol{\beta}, \boldsymbol{\alpha})/\partial\boldsymbol{\beta}^{\mathrm{t}}$ in (6) by a zero matrix. Having done so, the first set of

estimating equations in (6) can be written

$$K^{-1}\sum_{k=1}^{K}\left\{\left(\partial\boldsymbol{\mu}_k^{t}/\partial\boldsymbol{\beta}\right)\left(V_k^{-1}\right)_{11}\left(\mathbf{y}_k-\boldsymbol{\mu}_k\right)+\left(\partial\boldsymbol{\mu}_k^{t}/\partial\boldsymbol{\beta}\right)\left(V_k^{-1}\right)_{12}\left(\mathbf{s}_k-\boldsymbol{\sigma}_k\right)\right\}=\mathbf{0},$$

so that it will generally be necessary for V_k to be block-diagonal—that is, $V_{k12}\equiv 0$—for these equations to be unbiased regardless of whether $\mathrm{E}(\mathbf{s}_k)=\boldsymbol{\sigma}_k$, $k=1,\ldots,K$. Hence estimating equations of the type suggested in Section 2 seem natural to ensure mean parameter consistency under covariance misspecification. Candidate working specifications of V_{k22} are as given in (i)–(iii) above, in conjunction with $V_{k12}\equiv 0$. Intuitively the relative "information" about $\boldsymbol{\beta}$ from the covariance model and from the covariance between the response and its empirical covariance vector seems likely to be often small. Additional work will be required to determine circumstances in which potential efficiency gains in $\boldsymbol{\beta}$ estimation merit risking the possibility of some asymptotic bias.

4. Illustration

Consider now a clinical trial to study the effects of an anticonvulsant drug on neuropsychological function among patients with head injury. Data on 200 head injury patients were available (Temkin, 1989), 101 of whom had been randomly assigned to the anticonvulsant drug phenytoin, while the remainder were assigned to placebo. At the end of 1 year each patient's neuropsychological function was measured by performance intelligence quotient (IQ). Whether the patient had experienced one or more seizures during the preceding year was also recorded. At baseline each patient's injury severity was measured using several instruments, including a Glascow Coma Scale (GCS).

The effect of the anticonvulsant drug on the distribution of IQ and seizure occurrence was of primary interest in this trial. A binary seizure indicator was used rather than a count of the number of seizures since only 20 drug-treated and 16 placebo-treated patients experienced seizures. Since the mean determines the variance for a binary variate, the above estimating equations need to be reduced by eliminating the squared term in seizure occurrence, prior to application to this data set. Zhao and Prentice (1990) provide corresponding estimating equation expressions with binary response data.

Table 1 gives results of several analyses of this data set using the exponential quadratic score equations (6). In each analysis the mean IQ was allowed to depend linearly on treatment and GCS, while seizure occurrence was assumed to depend on these factors according to a binary logistic model; that is,

$$\mu_{k1}=\mathbf{x}_k^{t}\boldsymbol{\beta}_1,\quad \mu_{k2}=\exp\left(\mathbf{x}_k^{t}\boldsymbol{\beta}_2\right)\left\{1+\exp\left(\mathbf{x}_k^{t}\boldsymbol{\beta}_2\right)\right\}^{-1},$$

where $\mathbf{x}_k$ consists of a constant ($x_{k1}\equiv 1$), a treatment indicator (x_{k2}), and the GCS (x_{k3}) for the kth patient, $k=1,\ldots,200$. The variance of IQ was modelled using an exponential link as

$$\sigma_{k11}=\exp\left(\mathbf{x}_k^{t}\boldsymbol{\alpha}_{11}\right).$$

Finally, the covariance between IQ and seizure occurrence was modelled according to

$$\sigma_{k12}=\left(\mathbf{x}_k^{t}\boldsymbol{\alpha}_{12}\right)\left(\sigma_{k11}\sigma_{k22}\right)^{1/2},$$

where $\sigma_{k22}=\mu_{k1}(1-\mu_{k1})$, so that $\boldsymbol{\alpha}_{12}$ allows the magnitude of the correlation between IQ and seizure occurrence to depend on treatment and GCS. Note that the regression vector could be allowed to vary among the models for μ_{k1}, μ_{k2}, σ_{k11}, and σ_{k12}.

The first analysis of Table 1 applied the estimating equations (6) using the above independence working models for V_{k12} and V_{k22}. Corresponding to each regression parameter estimate both

Table 1

Regression analysis of neuropsychological responses in relation to anticonvulsant treatment and Glascow Coma Score (GCS) *using the exponential quadratic model scores* (6)

		Working variance matrices (V_{k12} and V_{k22})								
		Independence			Gaussian			Gaussian and higher correlations		
Regression variate	Codes/Ranges	Coeff[a]	SE(1)	SE(2)	Coeff	SE(1)	SE(2)	Coeff	SE(1)	SE(2)
IQ mean ($\hat{\beta}_1$)										
Constant	1	74.9	4.05	4.02	74.9	4.05	4.02	74.4	4.29	3.96
Treatment	0:placebo, 1:Rx	−.173	2.61	2.60	−.136	2.61	2.60	−.055	2.93	2.59
GCS	1–10	2.05	.344	.334	2.05	.344	.334	2.09	.365	.330
Seizure occurrence mean ($\hat{\beta}_2$)										
Constant	1	−.646	.559	.497	−.650	.558	.497	−.618	.604	.494
Treatment	0:placebo, 1:Rx	.291	.364	.375	.293	.364	.375	.255	.403	.371
GCS	1–10	−.109	.051	.048	−.108	.051	.048	−.110	.054	.048
IQ variance ($\hat{\alpha}_{11}$)										
Constant	1	6.42	.225	.285	6.42	.218	.285	6.39	.246	.274
Treatment	0:placebo, 1:Rx	−.271	.195	.200	−.268	.195	.200	−.248	.216	.192
GCS	1–10	−.046	.026	.025	−.046	.025	.025	−.043	.027	.024
IQ and seizure occurrence correlation ($\hat{\alpha}_{12}$)										
Constant		−.188	.218	.205	−.175	.215	.201	−.179	.235	.206
Treatment	0:placebo, 1:Rx	−.148	.133	.144	−.143	.133	.142	−.161	.165	.145
GCS	1–10	.008	.019	.018	.007	.018	.019	.008	.021	.018
Logistic transform of IQ and seizure correlation ($\hat{\alpha}_{12}$)										
Constant	1	−.380	.450	.426	−.357	.444	.418	−.360	.483	.429
Treatment	0:placebo, 1:Rx	−.308	.280	.301	−.298	.280	.295	−.331	.345	.304
GCS	1–10	.017	.040	.038	.015	.039	.037	.017	.043	.038

[a]Coeff refers to the estimated regression coefficient estimate, SE(1) to the corresponding empirical standard error estimate using (3), and SE(2) to the model-based standard error estimate using (8).

empirical (3) and model-based standard error estimates are given. Patients with high versus low values of GCS evidently have higher mean IQ, lower seizure probability, and possibly reduced IQ variance. The correlation between IQ and seizure rate does not appear to relate to GCS. In comparison, treatment did not appear to affect the IQ mean or variance, the seizure rate, or the correlation between IQ and seizure occurrence. Simultaneous tests of no treatment effect on the response means, or the response means and covariances, are also readily conducted. For example, an asymptotic χ_2^2 statistic for a hypothesis of no treatment effect on IQ or seizure occurrence takes value

$$(-.173 \quad .291)\begin{pmatrix} 2.61^2 & -.158 \\ -.158 & .364^2 \end{pmatrix}^{-1}\begin{pmatrix} -.173 \\ .291 \end{pmatrix} = .65,$$

using (3). The empirical and model-based standard error estimates are in reasonable agreement for both mean and covariance parameters.

The parameters in the covariance model σ_{k12} are restricted in that $-1 \leqslant \mathbf{x}_k^t \boldsymbol{\alpha}_{12} \leqslant 1$. Hence one may hope to improve both the properties of the iterative procedure (9) and the asymptotic distributional approximations by considering an alternate parametrization that forces the estimated correlations to fall within $(-1, 1)$. The final rows of Table 1 provide estimates of $\boldsymbol{\alpha}_{12}$ using the "logistic" transform

$$\sigma_{k12} = \{\exp(\mathbf{x}_k^t \boldsymbol{\alpha}_{12}) - 1\}\{\exp(\mathbf{x}_k^t \boldsymbol{\alpha}_{12}) + 1\}^{-1}(\sigma_{k11}\sigma_{k12})^{1/2}.$$

The estimates of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\alpha}_{11}$ and their estimated standard errors were virtually unchanged by this reparametrization and hence are not given. The estimated correlations also agree closely under the two parametrizations. For example, the estimated correlation between IQ and seizure occurrence for a treated patient with a GCS of 5 is $-.296$ under the linear correlation model, as compared to $-.293$ under the logistic correlation model.

The second analysis of Table 1 uses the same mean and covariance models in conjunction with Gaussian working variance matrices. The estimates of all parameters, as well as the estimates of α_{12} under the above logistic model, agree closely with those under the independence working model for V_{k12} and V_{k22}. Furthermore, both the empirical and model-based standard error estimates are virtually identical under the two working variance specifications.

The final analysis of Table 1 adds third- and fourth-order correlations to the Gaussian forms for V_{k12} and V_{k22}, respectively. Estimates of the additional parameters are as follows: $\hat{\gamma}_{111} = -.550$ (.260), $\hat{\gamma}_{112} = -.523$ (.122), $\hat{\gamma}_{122} = -.0399$ (.077), $\hat{\delta}_{1111} = -.142$ (.633), $\hat{\delta}_{1112} = .599$ (.295), and $\hat{\delta}_{1122} = -.052$ (.128), where empirical standard error estimates, calculated by ignoring the random variation in $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$, are given in parentheses. In spite of the fact that certain γ values are evidently nonzero, the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ and their standard error estimates differ little from the preceding analyses. The empirical standard error estimates tend to be somewhat larger than the model-based estimates and somewhat larger than the empirical standard error estimates using independence and Gaussian working variance matrices. There is little to suggest that the additional complexity of inserting estimates of third- and fourth-order correlations has increased estimating efficiency in this application.

Table 2 presents corresponding analyses using the simple estimating equations (1) and (2). More specifically, the analyses of Table 2 repeat those of Table 1 while setting $D_{k21} \equiv 0$ and $V_{k12} \equiv 0$ so that mean parameter estimation retains consistency under covariance model misspecification. The parameter estimates and both standard error estimates agree very closely with the corresponding estimates in Table 1. There would seem to be little reason to risk asymptotic bias in the mean parameter estimates through the use of (6), rather than the simpler (1) and (2), in this application. The final analysis of Table 2 uses the empirical estimates

Table 2

Regression analysis of neuropsychological responses in relation to anticonvulsant treatment and Glascow Coma Score (GCS) *using the simple estimating equations* (1) *and* (2)

		Working variance matrices ($V_{k12} \equiv 0$ and V_{k22})								
		Independence			Gaussian			Gaussian and higher correlations		
Regression variate	Codes/Ranges	Coeff[a]	SE(1)	SE(2)	Coeff	SE(1)	SE(2)	Coeff	SE(1)	SE(2)
IQ mean ($\hat{\beta}_1$)										
Constant	1	74.9	4.05	4.02	74.8	4.05	4.02	74.8	4.05	4.02
Treatment	0:placebo, 1:Rx	−.174	2.60	2.61	−.129	2.61	2.60	−.126	2.61	2.60
GCS	1–10	2.05	.343	.334	2.05	.343	.334	2.05	.343	.334
Seizure occurrence mean ($\hat{\beta}_2$)										
Constant	1	−.642	.556	.497	−.641	.556	.497	−.641	.556	.497
Treatment	0:placebo, 1:Rx	.290	.364	.375	.292	.364	.375	.292	.364	.375
GCS	1–10	−.109	.051	.048	−.109	.051	.048	−.109	.051	.048
IQ variance ($\hat{\alpha}_{11}$)										
Constant	1	6.42	.221	.285	6.42	.214	.285	6.41	.216	.274
Treatment	0:placebo, 1:Rx	−2.70	.195	.200	−.267	.195	.200	−.269	.206	.193
GCS	1–10	−.046	.026	.025	−.047	.024	.025	−.046	.025	.024
IQ and seizure occurrence correlation ($\hat{\alpha}_{12}$)										
Constant	1	−.188	.232	.202	−.174	.227	.205	−.171	.215	.205
Treatment	0:placebo, 1:Rx	−.148	.142	.142	−.144	.143	.144	−.144	.146	.145
GCS	1–10	.008	.021	.018	.007	.020	.018	.007	.019	.018
Logistic transform of IQ and seizure correlation ($\hat{\alpha}_{12}$)										
Constant	1	−.381	.477	.419	−.355	.468	.425	−.346	.437	.424
Treatment	0:placebo, 1:Rx	−.308	.300	.295	−.299	.302	.300	−.298	.304	.301
GCS	1–10	.017	.042	.037	.014	.041	.038	.014	.039	.038

[a]Coeff refers to the estimated regression coefficient estimate, SE(1) to the corresponding empirical standard error estimate using (3), and SE(2) to the model-based standard error estimate using (8).

$\hat{\delta}_{1111} = -.140$ (.641), $\hat{\delta}_{1112} = .549$ (.297), and $\hat{\delta}_{1122} = -.031$ (.129) in the working variance matrices V_{k22}.

5. Discussion

Expression (5) provides a class of models for a general multivariate response vector that is parametrized in terms of the mean and covariance of the response, and a shape function. Expression (6) gives a corresponding class of score estimating equations for the mean and covariance parameters that depend on the shape function only in terms of the third and fourth central moments of the response. Univariate response quadratic estimating equations of this type have been considered by Crowder (1987). Solving the score equations (6) will provide consistent mean and covariance parameter estimates for a range of working specifications for the third and fourth moments. Special cases include independence and Gaussian specifications as well as a relaxation of Gaussian moments that includes empirical estimates of third- and fourth-order correlations. It may be that third and fourth moment specifications under a simple independence model will lead to estimators of mean and covariance parameters of acceptable efficiency, while the use of Gaussian moment specifications with empirical third- and fourth-order correlations may lead to a procedure of high asymptotic efficiency over a range of sampling distributions. Analytic and numerical studies will be necessary to establish such efficiency properties, and to examine small-sample properties of the mean and covariance parameter estimators and of the scores on the left side of (6). See Godambe and Thompson (1989) for a recent discussion of optimality criteria for estimating functions.

It may also be possible to develop modifications to the estimation procedure that will improve small-sample distributional approximations—for example, by correcting covariance parameter estimates to acknowledge the estimation of mean parameters. See McCullagh and Nelder (1989, Chap. 10) for a discussion of related topics for mean and variance parameter estimation with a univariate response. They consider estimating equations that are a special case of (6), as well as estimating equations based on the deviance from Nelder and Pregibon's (1987) extended quasilikelihood, the latter of which need not be unbiased for variance parameters (Davidian and Carroll, 1988).

Upon inserting $D_{k21} \equiv 0$, $V_{k12} \equiv 0$ in the score equations (6), one obtains a special case of the ad hoc estimating equations (1) and (2) that yield consistent estimators of mean parameters even if the covariance model is misspecified. In most situations there would seem to be little useful information about the mean parameter β from either the covariance model or from the dependence between the response vector and its empirical covariance vector, at least if the covariance model is fairly flexible, so that setting $D_{k21} \equiv 0$, $V_{k21} \equiv 0$ would be expected to involve little efficiency loss. Again, further analytic and numerical work will be required to confirm this expectation.

The methods discussed above provide a practical means of analyzing multivariate response data, even if the response vectors are of large and variable dimension. The emphasis on marginal parameters is conducive to meaningful modelling in such circumstances, and numerical aspects of the use of (9) are straightforward under certain choices for the working variance matrices V_k. For example, under an independence specification for V_k, only the response variance V_{k11}, of dimension n_k, need be inverted numerically, and such numerical inversion may be avoided for various "patterned" covariance matrices. Additional study of the stability and convergence properties of (9), especially under constrained covariance model parameters (α), would be worthwhile. In the preceding illustration the number of iterations from common starting values agreed closely under the linear and logistic correlation models.

The ability of the class of estimating equations (6) to encompass a broad range of univariate and multivariate data analysis problems indicates that efforts to address the theoretical and practical issues mentioned above will be quite worthwhile.

ACKNOWLEDGEMENTS

This work was supported by Grant CA-53996 awarded by the National Institutes of Health. The authors would like to thank Dr Nancy Temkin for access to the data used in Section 4.

RÉSUMÉ

On introduit des équations d'estimation généralisée d'une manière ad hoc pour la matrice de covariance de réponses multidimensionnelles. Ces équations doivent être résolues simultanément avec les équations des scores pour un modèle linéaire généralisé relatif aux paramètres des moyennes. On utilise une classe de modèles exponentiels quadratiques pour les équations simultanées des paramètres relatifs aux paramètres de moyenne et de covariance, et on fait des propositions pour l'emploi de telles équations. On donne quelques commentaires sur les mérites relatifs des approches 'ad hoc' et par un modèle pour l'estimation, et on l'illustre par une régression d'une réponse bidimensionnelle.

REFERENCES

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis. Theory and Practice*. Cambridge, Massachusetts: MIT Press.

Crowder, M. (1985). Gaussian estimation for correlated binomial data. *Journal of the Royal Statistical Society, Series B* **47**, 229–237.

Crowder, M. (1987). On linear and quadratic estimating functions. *Biometrika* **74**, 591–597.

Davidian, M. and Carroll, R. J. (1988). A note on extended quasi-likelihood. *Journal of the Royal Statistical Society, Series B* **50**, 74–82.

Frydenberg, M. and Lauritzen, S. L. (1989). Decomposition of maximum likelihood in mixed graphical interactive models. *Biometrika* **76**, 539–555.

Godambe, V. P. and Thompson, M. E. (1989). An extension of quasi-likelihood estimation (with Discussion). *Journal of Statistical Planning and Inference* **22**, 137–172.

Gourieroux, C., Montfort, A., and Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica* **52**, 681–700.

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics* **17**, 31–57.

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

Manski, C. and McFadden, D. (1981). *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, Massachusetts: MIT Press.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, 2nd edition. London: Chapman and Hall.

Nelder, J. A. and Pregibon, D. (1987). An extended quasilikelihood function. *Biometrika* **74**, 221–232.

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

Temkin, N. R. (1989). Does phenytoin prevent posttraumatic seizures? One year follow-up results of a randomized double blind study in 407 patients (Abstract). *Journal of Neurosurgery* **70**, 314–315.

Whittle, P. (1961). Gaussian estimation in stationary time series. *Bulletin of the International Statistical Institute* **39**, 1–26.

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

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

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

Received December 1989; *revised July* 1990; *accepted August* 1990.

APPENDIX 1

Asymptotic Normality of $\{K^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t, K^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})^t\}$

As above suppose that $\boldsymbol{\mu}_k = \boldsymbol{\mu}_k(\boldsymbol{\beta})$ and $\boldsymbol{\sigma}_k = \boldsymbol{\sigma}_k(\boldsymbol{\beta}, \boldsymbol{\alpha})$, $k = 1, \ldots, K$. Let $\boldsymbol{\theta}^t = (\boldsymbol{\beta}^t, \boldsymbol{\alpha}^t)$ and let $\hat{\boldsymbol{\theta}}^t = (\hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\alpha}}^t)$ solve the estimating equations of the form (6) with weight matrices $V_k = V_k\{\boldsymbol{\theta}, \hat{\boldsymbol{\gamma}}(\boldsymbol{\theta})\}$, where $\hat{\boldsymbol{\gamma}}(\boldsymbol{\theta})$ is a $K^{1/2}$-consistent estimate of a parameter $\boldsymbol{\gamma}$ that characterizes V_{k12} and V_{k22}. Development of the

asymptotic distribution of $K^{1/2}(\hat{\theta} - \theta)$ then follows the appendix of Liang and Zeger (1986), essentially without change. Specifically, under regularity conditions one can approximate $K^{1/2}(\hat{\theta} - \theta)$ by

$$\left[-K^{-1}\sum_k \delta \mathbf{U}_k\{\theta, \hat{\gamma}(\theta)\}/\delta\theta\right]^{-1}\left[K^{-1/2}\sum_k \mathbf{U}_k\{\theta, \hat{\gamma}(\theta)\}\right], \tag{A.1}$$

where $\mathbf{U}_k\{\theta, \hat{\gamma}(\theta)\} = \sum_k D_k^t V_k^{-1}\mathbf{f}_k$, with all quantities evaluated at $\{\theta, \hat{\gamma}(\theta)\}$ and

$$\delta \mathbf{U}_k\{\theta, \hat{\gamma}(\theta)\}/\delta\theta = \partial \mathbf{U}_k\{\theta, \hat{\gamma}(\theta)\}/\partial\theta + \left[\partial \mathbf{U}_k\{\theta, \hat{\gamma}(\theta)\}/\partial\hat{\gamma}\right]\left[\partial\hat{\gamma}(\theta)/\partial\theta\right].$$

Taylor expansion at fixed θ gives

$$K^{-1/2}\sum \mathbf{U}_k\{\theta, \hat{\gamma}(\theta)\} = K^{-1/2}\sum \mathbf{U}_k\{\theta, \gamma\} + \{K^{-1}\sum \mathbf{U}_k(\theta, \gamma)/\partial\gamma\}K^{1/2}\{\hat{\gamma}(\theta) - \gamma\} + o_p(1). \tag{A.2}$$

Now $\partial U_k(\theta, \gamma)/\partial\gamma$ is a linear function of $\mathbf{f}_k$ that has mean zero, $k = 1, \ldots, K$, so that $K^{-1}\sum \partial U_k(\theta, \gamma)/\partial\gamma$ is quite generally $o_p(1)$, whereas $K^{1/2}\{\hat{\gamma}(\theta) - \gamma\}$ is $o_p(1)$ by assumption. Hence the asymptotic distribution of the score statistic on the left side of (A.2) is identical to that for the first term on the right side, which quite generally has an asymptotic normal distribution with mean zero and covariance matrix

$$\lim_{K\to\infty}\left\{K^{-1}\sum_{k=1}^{K} D_k^t V_k^{-1}\operatorname{cov}\mathbf{y}_k V_k^{-1} D_k\right\}. \tag{A.3}$$

The desired asymptotic result for $K^{1/2}(\hat{\theta} - \theta)$ then follows from noting that the matrix in (A.1) is consistent for $K^{-1}\sum D_k^t V_k^{-1} D_k$. This result follows from the fact that quite generally $K^{-1}\sum \partial U_k\{\theta, \hat{\gamma}(\theta)\}/\partial\hat{\gamma}$ is $o_p(1)$ and $\partial\hat{\gamma}(\theta)/\partial\theta$ is $o_p(1)$ so that $[-K^{-1}\sum \delta \mathbf{U}_k\{\theta, \hat{\gamma}(\theta)\}/\delta\theta]$ and $[-K^{-1}\sum \partial \mathbf{U}_k\{\theta, \hat{\gamma}(\theta)\}/\partial\theta]$ converge in probability to the same limit, while generally

$$-K^{-1}\sum \partial \mathbf{U}_k\{\theta, \hat{\gamma}(\theta)\}/\partial\theta = K^{-1}\sum D_k^t V_k^{-1} D_k + o_p(1),$$

since the terms collected in $o_p(1)$ are again sample averages of linear functions of the $\mathbf{f}_k$. Expression (3) arises by noting that (A.3) can generally be consistently estimated by inserting $\mathbf{f}_k\mathbf{f}_k^t$ as an estimator for $\operatorname{var}\mathbf{y}_k$.

Suppose now that one requires $D_{k21} \equiv 0$ and $V_{k12} \equiv 0$ in (6). The asymptotic normality of $K^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ with consistent variance estimator (4) then follows directly from the results of Liang and Zeger (1986) upon noting that $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta})$ solving (2) is quite generally a $K^{1/2}$-consistent estimator of $\boldsymbol{\alpha}$. Note that this result does not require the covariance form $\boldsymbol{\sigma}_k = \sigma_k(\boldsymbol{\beta}, \boldsymbol{\alpha})$ to have been correctly specified.

Appendix 2

Derivation of the Score Equations (6) *and Information Matrix* (8)

For ease of notation drop the subscript k in (5) and so that $\mathbf{y}^t = (y_1, \ldots, y_n)$ is distributed according to

$$\Pr(\mathbf{y}) = \Delta^{-1}\exp\{\mathbf{y}^t\theta + \mathbf{w}^t\boldsymbol{\lambda} + c(\mathbf{y})\}.$$

The marginal mean $\boldsymbol{\mu}$ and product moment $\boldsymbol{\eta} = (\eta_{11}, \eta_{12}, \ldots, \eta_{22}, \ldots)$, where $\eta_{ij} = \sigma_{ij} + \mu_i\mu_j$, are given by

$$\boldsymbol{\mu} = \sum_{\mathbf{y}} \mathbf{y}\exp\{\mathbf{y}^t\theta + \mathbf{w}^t\boldsymbol{\lambda} + c(\mathbf{y})\}\Delta^{-1}, \quad \boldsymbol{\eta} = \sum_{\mathbf{y}} \mathbf{w}\exp\{\mathbf{y}^t\theta + \mathbf{w}^t\boldsymbol{\lambda} + c(\mathbf{y})\}\Delta^{-1}.$$

The inverse Jacobian of the transformation from canonical parameters $\theta, \boldsymbol{\lambda}$ to marginal parameters $\boldsymbol{\mu}, \boldsymbol{\eta}$ is

$$\tilde{V} = \begin{pmatrix} \partial\boldsymbol{\mu}^t/\partial\theta & \partial\boldsymbol{\eta}^t/\partial\theta \\ \partial\boldsymbol{\mu}^t/\partial\boldsymbol{\lambda} & \partial\boldsymbol{\eta}^t/\partial\boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \operatorname{var}\mathbf{y} & \operatorname{cov}(\mathbf{y}, \mathbf{w}) \\ \operatorname{cov}(\mathbf{w}, \mathbf{y}) & \operatorname{var}\mathbf{w} \end{pmatrix} = \begin{pmatrix} \tilde{V}_{11} & \tilde{V}_{12} \\ \tilde{V}_{21} & \tilde{V}_{22} \end{pmatrix}.$$

Since $\tilde{V}$ is simply the variance matrix for $(\mathbf{y}^t, \mathbf{w}^t)$, the transformation will be one-to-one unless the distribution of $(\mathbf{y}^t, \mathbf{w}^t)$ is degenerate. The log-likelihood contribution can be written

$$l = \mathbf{y}^t\boldsymbol{\theta} + \mathbf{w}^t\boldsymbol{\lambda} + c(\mathbf{y}) - \log \Delta.$$

Using the chain rule, the score statistic contribution from the observation $\mathbf{y}$ can now be written

$$\begin{pmatrix} \partial l/\partial \boldsymbol{\beta} \\ \partial l/\partial \boldsymbol{\alpha} \end{pmatrix} = \tilde{D}^t \tilde{V}^{-1} \tilde{\mathbf{f}},$$

where

$$\tilde{D} = \begin{pmatrix} \partial \boldsymbol{\mu}/\partial \boldsymbol{\beta}^t & 0 \\ \partial \boldsymbol{\eta}/\partial \boldsymbol{\beta}^t & \partial \boldsymbol{\eta}/\partial \boldsymbol{\alpha}^t \end{pmatrix} \quad \text{and} \quad \tilde{\mathbf{f}} = \begin{pmatrix} \mathbf{y} - \boldsymbol{\mu} \\ \mathbf{w} - \boldsymbol{\eta} \end{pmatrix}.$$

The corresponding information matrix contribution is $\tilde{D}^t \tilde{V}^{-1} \tilde{D}$.

Note that one can write $\tilde{\mathbf{f}} = U\mathbf{f}$, $\tilde{V} = UVU^t$ and $\tilde{D} = UD$, where $\mathbf{f}$, V, and D are, aside from the subscript k, as defined in Section 2. The block-diagonal matrix U is given explicitly by

$$U = \begin{pmatrix} I_n & 0 \\ C & I_{n(n+1)/2} \end{pmatrix},$$

where I_m denotes an identity matrix of dimension m and

$$C^t = \left(C_1^t \mid \cdots \mid C_n^t\right), \quad \text{where} \quad C_i = \left(0_i \mid X_{n-i+1} + \mu_i I_{n-i+1}\right),$$

where 0_i denotes a zero matrix of dimension $(n - i + 1) \times (i - 1)$, which is absent for $i = 1$, and X_{n-i+1} is a zero matrix aside from the first column, the transpose of which is $(\mu_i, \ldots, \mu_n)$. Substitution then allows the score and information contributions to be written $D^t V^{-1}\mathbf{f}$ and $D^t V^{-1} D$, respectively, as is required for expressions (6) and (8).