*Biometrika* (2000), **87**, 4, *pp.* 823–836


# Improving generalised estimating equations using quadratic inference functions

By ANNIE QU

*Department of Statistics, Oregon State University, Corvallis, Oregon 97331, U.S.A.*

qu@stat.orst.edu

BRUCE G. LINDSAY and BING LI

*Department of Statistics, 326 Thomas Building, The Pennsylvania State University, University Park, Pennsylvania 16802, U.S.A.*

bgl@psu.edu   bing@stat.psu.edu

## Summary

Generalised estimating equations enable one to estimate regression parameters consistently in longitudinal data analysis even when the correlation structure is misspecified. However, under such misspecification, the estimator of the regression parameter can be inefficient. In this paper we introduce a method of quadratic inference functions that does not involve direct estimation of the correlation parameter, and that remains optimal even if the working correlation structure is misspecified. The idea is to represent the inverse of the working correlation matrix by the linear combination of basis matrices, a representation that is valid for the working correlations most commonly used. Both asymptotic theory and simulation show that under misspecified working assumptions these estimators are more efficient than estimators from generalised estimating equations. This approach also provides a chi-squared inference function for testing nested models and a chi-squared regression misspecification test. Furthermore, the test statistic follows a chi-squared distribution asymptotically whether or not the working correlation structure is correctly specified.

*Some key words*: Generalised estimating equation; Generalised method of moments; Linear approximate inverse; Longitudinal data; Quadratic inference function; Quasilikelihood.

## 1. Introduction

Generalised estimating equations were developed from generalised linear models (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989) and quasilikelihood (Wedderburn, 1974; McCullagh, 1983) to deal with nonnormal correlated longitudinal data. Liang & Zeger (1986) introduced the ingenious idea of using a working correlation matrix with a small set of nuisance parameters $\alpha$ to avoid specification of correlation between measurements within the cluster. The generalised estimating equation estimators of the regression parameter $\beta$ are consistent even when the true correlation matrix is not an element of the class of working correlation matrices, and are efficient when the working correlation is correctly specified, in the sense that the asymptotic variance of $\hat{\beta}$ reaches a Cramér–Rao-type lower bound.

When the working correlation is misspecified, however, the moment estimator of the nuisance parameter $\alpha$ suggested by Liang & Zeger (1986) no longer results in the optimal estimation of $\beta$. Furthermore, their estimator of $\alpha$ does not exist in some simple cases of misspecification (Crowder, 1995).

The purpose of this paper is to introduce a different strategy for estimating the working correlation so that the estimator always exists, and, even if the correlation is misspecified, the regression estimator remains optimal within the assumed family, and hence is more efficient than Liang & Zeger's regression estimator under the same misspecification. To motivate our method, consider the simple case where $\beta$ is a scalar. The asymptotic variance of the estimator of $\beta$, $\sigma^2(\alpha)$, say, is a function of $\alpha$, the working correlation parameters. If, instead of estimating $\alpha$ by the method of moments, we choose $\hat{\alpha}$ by minimising $\sigma^2(\alpha)$ among all possible $\alpha$, then the existence of $\hat{\alpha}$ would be guaranteed, and furthermore the estimator of $\beta$ would be optimal among the choices of $\alpha$, whether or not the working correlation structure is correctly specified. Since we minimise the empirical asymptotic variance rather than the parametric one, we require no additional moment assumption than does the generalised estimating equation method. This idea of maximising the empirical information was introduced by Lindsay (1985) to construct optimally weighted conditional scores free of nuisance parameters, and was also used in unpublished work by B. Li to derive nonparametric optimal estimating equations for independent errors without assuming a functional mean-variance relationship.

If $\beta$ is a scalar, then the above minimisation is straightforward. However, if $\beta$ is a vector, we have to minimise an empirical asymptotic covariance matrix, which may not have a Löwner-optimal solution for a typical problem. To circumvent this problem, we introduce a quadratic inference function method based on the generalised method of moments (Hansen, 1982). This enables us to embed the multivariate working correlation problem into a larger linear optimisation problem, where the Löwner-optimal solution exists and is explicit.

A quadratic inference function has the form $Q(\beta) = g'C^{-1}g$, where $g$ is a set of estimating functions based on moment assumptions and $C$ is the estimated variance of $g$. The quadratic inference function plays an inferential role similar to that of the negative of the loglikelihood, with parallel construction of point estimators and chi-squared tests. The associated point estimator is the minimiser of $Q(\beta)$, and has the minimum asymptotic variance matrix, in the Löwner ordering, over all estimating functions constructed by linear combinations of the elements of $g$ (Hansen, 1982; Lindsay, 1982). Our simulation results will show that the quadratic inference function method with appropriate scores $g$ is more efficient than the generalised estimating equation approach when the working structure is misspecified.

For hypothesis testing, we establish some new inferential properties for the $Q(\beta)$-based test statistics that extend the results of Hansen (1982) and Lee (1996). The test statistics we propose follow a $\chi^2$ distribution asymptotically whether or not the working correlation structure is correctly specified; this contrasts with Rotnitzky & Jewell's (1990) score test result in that their test distributions are not robust against misspecified working assumptions. The test statistics are shown to be asymptotically noncentral $\chi^2$ under local alternatives.

Another method for increasing efficiency was proposed by Prentice & Zhao (1991), who jointly solved estimating equations associated with the response mean and covariance matrix. However, this requires the functional form of the third and fourth moments, and so is more restrictive in assumptions than the generalised estimating equation itself. Our method requires no such additional assumption.

Section 2 introduces the quadratic inference function based on the generalised method of moments (Hansen, 1982) and the linear approximate inverse described in unpublished work by B. G. Lindsay, A. Qu and S. Lele. Section 3 discusses the inferential properties of quadratic inference functions for $\chi^2$ testing, and § 4 illustrates comparisons of the generalised estimating equation and extended quadratic inference function methods using biomedical data for longitudinal binary outcomes. The final section provides a brief discussion.

## 2. Quadratic inference functions

### 2·1. *Quasilikelihood equations and generalised estimating equations*

Let $y_{it}$ be an outcome variable and $x_{it}$ be a $q \times 1$ vector of covariates, observed at times $t = 1, \ldots, n_i$ for subjects $i = 1, \ldots, N$. We assume that the observations from different subjects are independent, but that those within the same subject are dependent. We assume further that $E(y_{it}) = \mu(x'_{it}\beta)$. We ask how $\beta$ can be most efficiently estimated using this information.

Given a $k$-dimensional score vector $m(y, x, \beta)$ that satisfies the moment assumption $E(m) = 0$, the estimating function $g$ that is the optimal linear combination of the elements of $m$ based on the projection theorem (Small & McLeish, 1994, p. 79) is

$$g_{\text{opt}} = (E\dot{m})'\Sigma^{-1}m, \tag{1}$$

where $\dot{m}$ is the $k \times q$ matrix whose entries are $\partial m_i/\partial\beta$, and $\Sigma$ is the $k \times k$ covariance matrix of $m$. The optimality is in the sense that the asymptotic variance of the solution to $g_{\text{opt}}(\beta) = 0$ reaches the minimum among all estimating equations formed by taking linear combinations of $m$.

To formulate our problem, let $y_i$ be the vector $(y_{i1}, \ldots, y_{in_i})'$, $\mu_i$ be $(\mu_{i1}, \ldots, \mu_{in_i})'$, $V_i$ be the covariance matrix of the vector $y_i$ and $\dot{\mu}_i$ be the $n_i \times q$ matrix

$$\{\partial\mu_{it}/\partial\beta : i = 1, \ldots, N;\, t = 1, \ldots, n_i\}.$$

Then, if we let $m$ be the vector $((y_1 - \mu_1)', \ldots, (y_N - \mu_N)')'$, the general formula (1) reduces to the quasilikelihood equation

$$g_{\text{opt}} = \sum \dot{\mu}'_i V_i^{-1}(y_i - \mu_i). \tag{2}$$

If $V_i$ is unknown, one might use (2) with empirical estimators $\hat{V}_i$ for the $V_i$. However, if the size of $V_i$ is large, there will be many nuisance parameter estimations, and a high risk of numerical error in the inversion of $\hat{V}_i$. To avoid this, Liang & Zeger (1986) introduced the idea of using a working correlation matrix $W(\alpha)$ which depends on fewer nuisance parameters $\alpha$. The common working correlation structure could be as simple as independent, equicorrelated, first-order autoregressive, AR(1), or could be unspecified. The use of (2) with estimated working parameters $\alpha$ is known as the method of generalised estimating equations.

### 2·2. *Quadratic inference functions*

We will model $R^{-1}$ by the class of matrices

$$\sum_{i=1}^{m} a_i M_i, \tag{3}$$

where $M_1, \ldots, M_m$ are known matrices and $a_1, \ldots, a_m$ are unknown constants. This is a sufficiently rich class that accommodates, or at least approximates, the correlation structures most commonly used. Note, however, that we do not need to assume that class (3) contains the true correlation matrix, as subsequent development does not depend on this assumption.

*Example* 1. Suppose $R(\alpha)$ is an equicorrelated matrix; it has 1's on the diagonal, and $\alpha$'s everywhere off the diagonal. Then $R^{-1}$ can be written as $a_0 M_0 + a_1 M_1$, where $M_0$ is the identity matrix and $M_1$ is a matrix with 0 on the diagonal and 1 off the diagonal. Here $a_0 = -\{(n-2)\alpha + 1\}/k_1$ and $a_1 = \alpha/k_1$, where $k_1 = (n-1)\alpha^2 - (n-2)\alpha - 1$ and $n$ is the dimension of $R$. Note that this is not a unique linear representation of $R^{-1}$; we could also choose $M_1$ to be the rank-1 matrix with 1 everywhere.

*Example* 2. Suppose $R(\alpha)$ is the first-order autoregressive correlation matrix, with $R_{ij} = \alpha^{|i-j|}$. The exact inversion of $R^{-1}$ can be written as a linear combination of three basis matrices; they are $M_0$, $M_1$ and $M_2$, where $M_0$ is the identity matrix, $M_1$ has 1 on the two main off-diagonals and 0 elsewhere, and $M_2$ has 1 on the corners $(1, 1)$ and $(n, n)$, and 0 elsewhere. Here $a_0 = (1 + \alpha^2)/k_2$, $a_1 = -\alpha/k_2$ and $a_2 = -\alpha^2/k_2$, where $k_2 = 1 - \alpha^2$. A simple approximation to $R^{-1}$ would use just $M_0$ and $M_1$. The third term of the inverse in this example captures the edge effect of the process AR(1) while the two-term approximation does not.

Although the above two cases will be our primary examples here, the pool of possibilities is considerably richer. The literature on multivariate normal models, for example, has considerable discussion of the situation where the covariance matrix $\Sigma$ has a parametric linear inverse structure of the type we describe. In this case there are complete and sufficient statistics for all parameters and sometimes explicit point estimators (Seely, 1971). In particular, we note the following important class of models, where the linear structure for $\Sigma^{-1}$ arises naturally from a linear structure for $\Sigma$.

Estimation of covariance matrices which are linear combinations, or whose inverses are linear combinations, of given matrices was intensively studied by Anderson (1969, 1970). Under the assumption of normality, if consistent estimators of the coefficients of the linear combinations are used to obtain the regression parameter, then the estimator of the regression parameters is asymptotically efficient (Anderson, 1973).

Consider the covariance matrices generated by balanced nested design structures. For example, suppose that the correlation matrix $R$ for a cluster of size 4 has the block structure $R = a_0 I + a_1 M_1 + a_2 M_2$, where $M_1$ has all entries equal to 1 and $M_2$ has the structure

$$M_2 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

This model might arise if we had two blocks of size 2 inside the cluster of size 4. Both $P_1 = 0{\cdot}25 M_1$ and $P_2 = 0{\cdot}5 M_2$ are projection matrices, corresponding to projection on to $(1, 1, 1, 1)'$ and span $\{(1, 1, 0, 0), (0, 0, 1, 1)\}$ respectively. Furthermore, the subspaces are nested, with $P_1 P_2 = P_2 P_1 = P_1$. We show that in this case $R^{-1}$ has the same linear structure.

Suppose we have a linear representation for $\Sigma$ of the form

$$a_0 I + a_1 P_1 + a_2 P_2 + \ldots + a_d P_d,$$

where the $P_j$ are projection matrices and there is closure under multiplication in the sense that, for every $(i, j)$, there exists $k$ such that $P_i P_j = P_k$, for some $k$; that is, the pairs of subspaces corresponding to the projections are either nested or orthogonal. Then we can write the inverse in the form $\Sigma^{-1} = b_0 I + b_1 P_1 + \ldots + b_d P_d$. The coefficients $b_j$ are determined by solving the equations generated by the relationship

$$(a_0 I + a_1 P_1 + a_2 P_2 + \ldots + a_d P_d)(b_0 I + b_1 P_1 + \ldots + b_d P_d) = 1I + 0P_1 + 0P_2 + \ldots + 0P_d.$$

This calculation is simplified if the $P_j$ are orthogonal and idempotent, as then we have $a_0 b_0 = 1$ and $a_0 b_j + a_j b_0 + a_j b_j = 0$ for every $j \geqslant 1$. If $a_0 = 1$, then $b_0 = 1$ and $b_j = -a_j/(1 + a_j)$ for all $j \geqslant 1$.

Substituting (3) into (2), consider the following class of estimating functions:

$$\sum_{i=1}^{N} \dot{\mu}_i' A_i^{-\frac{1}{2}} (a_1 M_1 + \ldots + a_m M_m) A_i^{-\frac{1}{2}} (y_i - \mu_i), \tag{4}$$

where $\dot{\mu}_i$ is the derivative of $\mu_i$ with respect to regression parameters $\beta$, and $A_i$ is the diagonal marginal covariance matrix for the $i$th cluster.

One approach to estimation would be to choose the parameters $a = (a_1, \ldots, a_m)$ so as to optimise some function of the information matrix associated with (4). Instead, we proceed as follows. Based on the form of the quasi-score, we define the 'extended score' $g_N$ to be

$$g_N(\beta) = \frac{1}{N} \sum_{i=1}^{N} g_i(\beta) = \frac{1}{N} \begin{pmatrix} \sum_{i=1}^{N} (\dot{\mu}_i)' A_i^{-\frac{1}{2}} M_1 A_i^{-\frac{1}{2}} (y_i - \mu_i) \\ \vdots \\ \sum_{i=1}^{N} (\dot{\mu}_i)' A_i^{-\frac{1}{2}} M_m A_i^{-\frac{1}{2}} (y_i - \mu_i) \end{pmatrix}. \tag{5}$$

Note that the estimating function (4) is a linear combination of elements of the extended score vector (5).

The vector $g_N$ contains more estimating equations than parameters, but these estimating equations can be combined optimally using the generalised method of moments (Hansen, 1982). This method is an extension of the minimum $\chi^2$ method introduced by Neyman (1949) and further developed by Ferguson (1958). The idea is to construct an estimator of $\beta$ by setting specified linear combinations of the $r$ estimating equations in $g_N$ as close to zero as possible when $r > q$. That is, $\hat{\beta}$ is obtained by minimising the weighted length of $g_N$:

$$\hat{\beta} = \arg\min_{\beta} g_N' W^{-1} g_N.$$

Hansen (1982) has shown that $\hat{\beta}$ is efficient if $W$ is the variance matrix of $g_N$. The intuition is that $W^{-1}$ gives less weight to the estimating equations with larger variances.

Based on the extended scores $g_N$, we define the quadratic inference function to be

$$Q_N(\beta) = g_N' C_N^{-1} g_N, \tag{6}$$

where $C_N = (1/N^2) \sum_{i=1}^{N} g_i(\beta) g_i'(\beta)$. Note that $C_N$ depends on $\beta$. Clearly, $Q_N$ is analogous to Rao's (1947) score test statistic and possesses inferential properties similar to the score test, but differs in that the dimension of the score $g_N$ is greater than that of $\beta$.

An example of $Q_N(\beta)$ is plotted in Fig. 1. There we have used the identity link

$$\mu(x_{it}, \beta) = x_{it}'\beta,$$

where $x_{it}' = (x_{it}^1, x_{it}^2)$, $\beta = (\beta_1, \beta_2)'$, for $i = 1, \ldots, 20$ and $t = 1, \ldots, 10$. The covariates $x_i^1$ and $x_i^2$ are generated independently from a multivariate normal distribution with mean $(0{\cdot}1, 0{\cdot}2, \ldots, 1{\cdot}0)$ and covariance matrix $I$. The response variable is defined by

$$y_i = \beta_1 x_i^1 + \beta_2 x_i^2 + \varepsilon_i,$$

where $\beta_1 = \beta_2 = 1$ and $\varepsilon_i$ is generated from a 10-dimensional normal distribution with mean 0, marginal variance 1 and an AR(1) correlation structure with autocorrelation $\alpha = 0{\cdot}7$. We construct the extended score $g_N$ using $M_0, M_1$ as in Example 1. Note that in this case $Q_N$ has a unique minimum point.

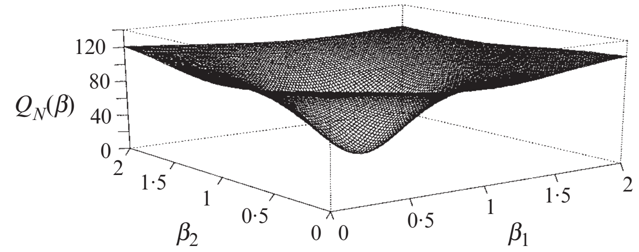


Fig. 1. Test statistic $Q_N(\beta) = g_N' C_N^{-1} g_N$ for two-dimensional $\beta$, where $g_N$ is defined by (5).

The quadratic inference function estimator $\hat{\beta}$ is then defined to be

$$\hat{\beta} = \arg\min_{\beta} Q_N(\beta). \tag{7}$$

The corresponding estimating equation for $\beta$ is

$$\dot{Q}_N(\beta) = 2\dot{g}_N' C_N^{-1} g_N - g_N' C_N^{-1} \dot{C}_N C_N^{-1} g_N = 0, \tag{8}$$

where $\dot{g}_N$ is the $mq \times q$ matrix $\{\partial g_N / \partial \beta\}$, $\dot{C}_N$ is the three-dimensional array $(\partial C_N/\partial \beta_1, \ldots, \partial C_N/\partial \beta_q)$, and $g_N' C_N^{-1} \dot{C}_N C_N^{-1} g_N$ is a $q \times 1$ vector

$$\{g_N' C_N^{-1} (\partial C_N / \partial \beta_i) C_N^{-1} g_N : i = 1, \ldots, q\}.$$

To solve equation (8), we implement the Newton–Raphson algorithm, which requires the second derivative of $Q_N$ in $\beta$:

$$\ddot{Q}_N(\beta) = 2\dot{g}_N' C_N^{-1} \dot{g}_N + R_N,$$

where

$$R_N = 2\ddot{g}' C^{-1} g - 4\dot{g}_N' C_N^{-1} \dot{C}_N C_N^{-1} g_N + 2 g_N' C_N^{-1} C_N^{-1} \dot{C}_N C_N^{-1} g_N - g_N' C_N^{-1} \ddot{C}_N C_N^{-1} g_N.$$

Here $\ddot{C}_N$ is a four-dimensional array $\{\partial^2 C_N / \partial \beta_i \, \partial \beta_j : i, j = 1, \ldots, q\}$, and $g_N' C_N^{-1} \ddot{C}_N C_N^{-1} g_N$ is a $q \times q$ matrix $\{g_N' C_N^{-1} (\partial^2 C_N / \partial \beta_i \, \partial \beta_j) C_N^{-1} g_N : i, j = 1, \ldots, q\}$. Asymptotically $\ddot{Q}_N(\beta)$ can be approximated by $2\dot{g}_N' C_N^{-1} \dot{g}_N$ since $R_N$ is $o_p(1)$. The Newton–Raphson method then iterates the following relationship to convergence:

$$\hat{\beta}^{(j+1)} = \hat{\beta}^{(j)} - \ddot{Q}_N^{-1}(\hat{\beta}^{(j)}) \dot{Q}_N(\hat{\beta}^{(j)}).$$

The optimality of the quadratic inference function estimator is easily established. Note that the second term of equation (8) is $O_p(N^{-1})$, so that solving (8) is asymptotically equivalent to solving

$$\dot{g}_N' C_N^{-1} g_N = 0. \tag{9}$$

This equation is presented only for the convenience of the asymptotic analysis; for estimation of $\beta$ we still recommend the definition in (7), which, among other features, removes ambiguity when (9) has multiple roots. Since $\dot{g}_N$ is nonrandom, we have $E(\dot{g}_N) = \dot{g}_N$. The matrix $C_N$ converges to $E(C_N)$ in probability; the weight in (9) therefore converges in probability to the optimal weight $(E\dot{g}_N)'(EC_N)^{-1}$. By the projection theorem (Lindsay, 1982; Small & McLeish, 1994, p. 79), it can be verified that (9) is optimal among the class of estimating equations

$$\sum_{r=1}^{m} H_r \sum_{i=1}^{N} \dot{\mu}_i' A_i^{-\frac{1}{2}} M_r A_i^{-\frac{1}{2}} (y_i - \mu_i) = 0, \tag{10}$$

where $H_r$ $(r = 1, \ldots, m)$ are $q \times q$ arbitrary nonrandom matrices. Note that, if $H_r = a_r I$ for the identity matrix $I$, then the left-hand side of (10) becomes the same as (4).

Hence, if the inverse of the true correlation matrix $R^{-1}$ belongs to the class $\sum_{r=1}^{m} a_r M_r$, then (9) is fully efficient, that is, as efficient as the quasilikelihood (2); if not, (9) will still be optimal within the family (10).

The role played by the quadratic inference function here is to embed the smaller model (4) into the larger model (10), for which optimisation is easily achieved. In addition, equation (8), which differs from (9) only by an ignorable term, enables us to use the objective function $Q$ in conjunction with our algorithm, as equation (9) does not correspond to the minimum of any criterion; see Hansen (1982) for proofs of normality and optimality.

### 2·3. *Simulation results for point estimates*

We now compare the quadratic inference function and generalised estimating equation methods by simulation. We generate data by simulation as in § 2·2, using both the AR(1) and the equicorrelated correlation structures. The two methods are applied to each sample and the mean squared error of the estimators is estimated by averaging $(\hat{\beta}_1 - \beta_1)^2 + (\hat{\beta}_2 - \beta_2)^2$ over all samples. The simulated relative efficiency, SRE, is defined as

$$\text{SRE} = \frac{\text{mean squared error of the generalised estimating equation estimator}}{\text{mean squared error of quadratic inference function estimator}}. \tag{11}$$

Table 1 records SRE over a variety of working assumptions and shows that, if the working correlation structure is misspecified, the quadratic inference function approach is more efficient than the generalised estimating equation method. In particular, when the true correlation structure is AR(1) with autocorrelation $\alpha = 0{\cdot}7$, but the working assumption is equicorrelated, SRE $= 1{\cdot}34$; and SRE $= 2{\cdot}07$ when the true structure is equicorrelated with common correlation $\alpha = 0{\cdot}7$ and the working assumption is AR(1), with $M_0$, $M_1$ and $M_2$ in Example 2 as basis matrices.

On the other hand, when the working structure is correct, the two methods are almost equivalent, with SRE in the range 0·97–0·99. Since the generalised estimating equation estimators of the nuisance parameter are the maximum likelihood estimators when $\varepsilon$ is normal and the working assumption is correct, the generalised estimating equation method is optimal in that case.

Table 1. *Simulated relative efficiency*, SRE, *of* $\beta$ *as defined in* (11), *calculated from* 10 000 *simulations, for* $E(y_{it}) = x_{it}'\beta$, *where* $\beta = (1, 1)'$, *and when the true nuisance parameter is* $\rho = 0{\cdot}3$ *and* $\rho = 0{\cdot}7$

| True $R$ | $\rho$ | Working $R$ Equicorrelated | AR(1) |
|---|---|---|---|
| Equicorrelated | 0·3 | 0·99 | 1·20 |
| | 0·7 | 0·99 | 2·07 |
| AR(1) | 0·3 | 1·04 | 0·97 |
| | 0·7 | 1·34 | 0·98 |

We have also performed simulations for correlated Poisson data, again demonstrating the superiority of the quadratic inference function approach to the generalised estimating equation method under misspecified working structure.

## 3. Chi-squared tests

In this section, we give the asymptotic limiting distribution of the quadratic inference function under the null hypothesis and local alternatives.

Suppose that the regression parameter $\beta$ is partitioned into $(\psi, \lambda)$, where $\psi$ is a regression parameter of interest with dimension $p$, and $\lambda$ is a nuisance regression parameter with dimension $q-p$. As a special case, we also allow $p=q$, with $\beta=\psi$ and $\lambda$ being absent. For testing the hypothesis $H_0: \psi = \psi_0$, we propose using $Q(\psi_0, \tilde{\lambda}) - Q(\hat{\psi}, \hat{\lambda})$, where

$$\tilde{\lambda} = \arg\min_{\lambda} Q(\psi_0, \lambda), \quad (\hat{\psi}, \hat{\lambda}) = \arg\min_{(\psi,\lambda)} Q(\psi, \lambda). \tag{12}$$

We define a parametric family of local alternatives to $P_{\beta_0}$ to be a set of distributions $\{P_\beta\}$, indexed by $\beta$ in some neighbourhood of $\beta_0$, satisfying the zero-mean assumption locally, that is $E_{\beta_0} g(\beta_0) = 0$, and LeCam's local asymptotic normality conditions (Hall & Mathiason, 1990).

To simplify the notation, let

$$\frac{\partial Q_N}{\partial \psi} = \dot{Q}_\psi, \quad \frac{\partial Q_N}{\partial \lambda} = \dot{Q}_\lambda, \quad \frac{\partial^2 Q_N}{\partial \psi^2} = \ddot{Q}_{\psi\psi}, \quad \frac{\partial^2 Q_N}{\partial \psi\, \partial \lambda} = \ddot{Q}_{\psi\lambda}, \quad \frac{\partial^2 Q_N}{\partial \lambda^2} = \ddot{Q}_{\lambda\lambda}.$$

Write

$$d_0' \Sigma^{-1} d_0 = \begin{pmatrix} J_{\psi\psi} & J_{\psi\lambda} \\ J_{\lambda\psi} & J_{\lambda\lambda} \end{pmatrix}.$$

Note that, if $\psi$ and $\lambda$ converge in probability to $\psi_0$ and $\lambda_0$ respectively, then $\frac{1}{2}\ddot{Q}_{\psi\psi}(\psi, \lambda)$ and $\frac{1}{2}\ddot{Q}_{\psi\lambda}(\psi, \lambda)$ converge in probability to $J_{\psi\psi}$ and $J_{\psi\lambda}$ respectively.

Theorem 1. *Suppose that* $\psi$ *has dimension* $p$, *and all required regularity conditions are satisfied. Then, under the null hypothesis,* $Q_N(\psi, \tilde{\lambda}) - Q_N(\hat{\psi}, \hat{\lambda})$ *is asymptotically* $\chi^2_p$; *under the local alternative hypothesis* $H_\alpha: \psi = \psi_0 + N^{-\frac{1}{2}} h_\psi$ *and* $\lambda = \lambda_0 + N^{-\frac{1}{2}} h_\lambda$, $Q_N(\psi, \tilde{\lambda}) - Q_N(\hat{\psi}, \hat{\lambda})$ *is asymptotically noncentral* $\chi^2_p$ *with noncentrality parameter* $\delta_\psi = h_\psi'(J_{\psi\psi} - J_{\psi\lambda} J_{\lambda\lambda}^{-1} J_{\psi\lambda}) h_\psi$.

Theorem 1 is proved in the Appendix. In the special case where no nuisance parameter is present, $Q_N(\beta_0) - Q_N(\hat{\beta})$ is asymptotically $\chi_q^2$ under the null hypothesis; under the local alternative hypothesis $H_\alpha : \beta_N = \beta_0 + N^{-\frac{1}{2}}h$, $Q_N(\beta_0) - Q_N(\hat{\beta})$ is asymptotically noncentral $\chi_q^2(\delta)$, with the noncentrality parameter $\delta = h'd_0'\Sigma^{-1}d_0h$.

We can also construct a goodness-of-fit statistic to test the model assumption

$$H_0 : E\{g_N(\beta)\} = 0. \tag{13}$$

Since $\hat{\beta}$ is obtained by equating $q$ linear combinations of the $mq$ components of $g_N$ to zero, there remain $mq - q$ linear combinations of $g_N$ that should be close to zero under the above model assumption. On these grounds, it is natural to use $Q_N(\hat{\beta})$ as the test statistic. This test was called an 'over-identifying restriction' test by Hansen (1982).

THEOREM 2 (*Hansen*, 1982). *Suppose $g_N$ has dimension $r$ and $\beta$ has dimension $q$ with $q < r$. Then, under the model assumption* (13), *the asymptotic distribution of $Q_N(\hat{\beta})$ is $\chi^2$ with $r - q$ degrees of freedom.*

For small samples and symmetric data, our simulation results show that the tests in Theorems 1 and 2 are conservative relative to their nominal $\chi^2$ distributions. This occurs because we have used $C_N$ in place of the covariance matrix of $g_N$. Gine & Mason (1997) show that in the univariate case the 'uncentred' $t$-statistics $C_N^{-\frac{1}{2}}g_N$ for symmetric data follow a sub-Gaussian distribution; that is, the moment generating function is bounded from above by the normal moment generating function. This explains why statistics based on $Q_N$ have lighter tails than the $\chi^2$ distributions.

*Example* 3. To examine the finite sample null distributions, we use the same simulated data as in § 2·3. We assume an equicorrelated working correlation matrix $R$, and so two basis matrices, $I$ and $M_1$, see Example 1, can be used for the expansion of $R^{-1}$. Therefore, our moment conditions are specified by the following vector of length 4:

$$g_N(\beta) = \frac{1}{N}\begin{pmatrix} \sum_{i=1}^N (x_i^1)' A_i^{-1}(y_i - \mu_i) \\ \sum_{i=1}^N (x_i^2)' A_i^{-1}(y_i - \mu_i) \\ \sum_{i=1}^N (x_i^1)' A_i^{-\frac{1}{2}} M_1 A_i^{-\frac{1}{2}}(y_i - \mu_i) \\ \sum_{i=1}^N (x_i^2)' A_i^{-\frac{1}{2}} M_1 A_i^{-\frac{1}{2}}(y_i - \mu_i) \end{pmatrix}.$$

Figure 2 shows Q–Q plots of $Q_N(\beta) - Q_N(\hat{\beta})$ based on 10 000 simulated datasets for two different covariance structures, namely equicorrelated or AR(1). It is clear that the plots indicate proximity to the $\chi_2^2$ distribution, even though the number of clusters $N = 20$ is fairly small. The corresponding plot for $Q_N(\hat{\beta})$ is very similar, also indicating proximity to $\chi_2^2$. For the test of Theorem 1 we partitioned $\beta$ into $(\psi, \lambda)$. A Q–Q plot, qualitatively very similar to Fig. 2, shows that $Q_N(\psi, \tilde{\lambda}) - Q_N(\hat{\psi}, \hat{\lambda})$ follows a $\chi_1^2$ distribution approximately. As expected, the plots, for which Fig. 2 are typical, show that all these tests are conservative in the tails in the sense that using the nominal size-$\alpha$ $\chi^2$ critical value for small $\alpha$ would lead to a test of size less than $\alpha$.

The results of these tests demonstrate a simplicity of use compared to the methods of Rotnitzky & Jewell (1990). Since our tests are analogous to Rao's score test, for a fair comparison we contrast our test only with their generalised score test. Rotnitzky & Jewell's score test statistics also follow an asymptotic chi-squared distribution with $q$ degrees of freedom, but a major drawback is that this limiting $\chi^2$ distribution relies on correct specification of the working correlation; otherwise it will not be $\chi^2$. Hence a consistent

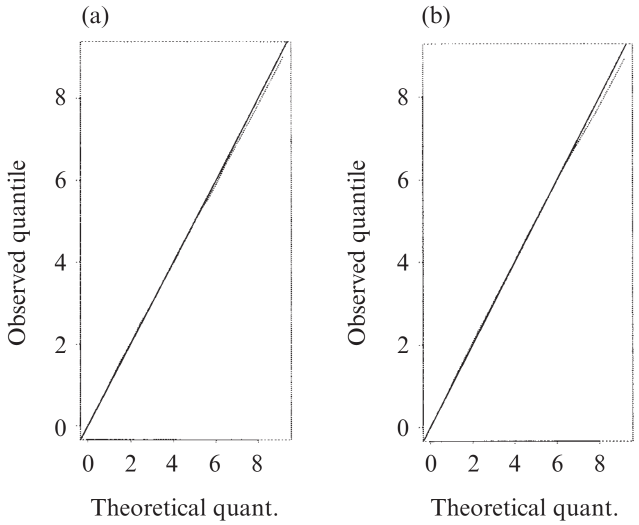


Fig. 2. Quantile–quantile plots of $Q_N(\beta) - Q_N(\hat{\beta})$ (dotted lines) relative to $\chi_2^2$ (solid lines) (a) when covariance structure is equicorrelated, (b) when covariance structure is AR(1).

estimator of $\text{cov}(y_i)$ is required. In contrast, our testing procedures do not have these limitations, because they follow $\chi^2$ distributions asymptotically regardless of the true correlation structure.

## 4. Application to longitudinal data for binary outcomes

In this section we analyse a longitudinal dataset from a Harvard University technical report by N. M. Laird, G. J. Beck and J. H. Ware.

This dataset is part of a study of the respiratory health effects of indoor and outdoor air pollution in six U.S. cities. One of the main issues of interest is the effect of maternal smoking on children's respiratory illness. In their report, Laird et al. used a random half-sample of the data collected on children in Steubenville, Ohio. The serial response variable for children from ages 7 to 10 is presented as a binary outcome with 0 or 1 denoting the absence or presence of respiratory illness. The maternal smoking habit is a dichotomous variable with 0 as yes and 1 as no. Laird et al. treated mothers' smoking habits as fixed at the status at the first visit. Also, the dataset was balanced by including only those children who had all four responses, at ages 7, 8, 9 and 10. Clearly, we would expect the measurements for each child to be serially correlated.

We apply both the generalised estimating equation and quadratic inference function methods to these data. The logistic link function is assumed for the marginal model, that is

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{ij}^{\text{A}} + \beta_2 x_{ij}^{\text{MS}} + \beta_3 x_{ij}^{\text{A}} x_{ij}^{\text{MS}},$$

where the covariates $x_{ij}$ are the intercept, the child's age A, the maternal smoking habit indicator MS and their interaction. Here $i$ denotes the $i$th child and $t$ denotes the $t$th measurement of the child. For the Bernoulli variable, the relationship between the marginal mean and variance is $A_{ij} = \mu_{ij}(1 - \mu_{ij})$. If we assume the working corelation to be $R_\alpha$, then the generalised estimating equation is

$$\sum_{i=1}^{N} \dot{\mu}_i' A_i^{-\frac{1}{2}} R_\alpha^{-1} A_i^{-\frac{1}{2}} (y_i - \mu_i) = 0, \tag{14}$$

where $\dot{\mu}_i = (\partial\mu_i/\partial\beta_0, \partial\mu_i/\partial\beta_1, \partial\mu_i/\partial\beta_2, \partial\mu_i/\partial\beta_3)'$ and $A_i = \mathrm{diag}(A_{ij})$. The solution of (14) with moment estimators for the working parameters is the generalised estimating equation estimator.

If we further assume the inverse of $R$ to be of the form $\alpha_1 I + \alpha_2 M_1$, where $M_1$ is as in Example 2 for the AR(1) structure, then the extended score is

$$g_N = \frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N} \dot{\mu}_i' A_i^{-1}(y_i - \mu_i) \\ \sum_{i=1}^{N} \dot{\mu}_i' A_i^{-\frac{1}{2}} M_1 A_i^{-\frac{1}{2}} (y_i - \mu_i) \end{pmatrix}.$$

The quadratic inference function estimator is then found by minimising, over $\beta$,

$$Q_N(\beta) = g_N' C_N^{-1} g_N,$$

where $C_N = (1/N^2)\sum g_i g_i'$.

Table 2 provides point estimates, standard errors and $t$-ratios using generalised estimating equations under independence, equicorrelated and AR(1) working correlations, and those using the quadratic inference function under AR(1) working correlation. Note that there is no theoretical difference between the generalised estimating equation and quadratic inference function methods under the equicorrelated working structure, since having an intercept in the regression model is confounded with the equicorrelation matrix for balanced data.

Table 2. *Comparison of generalised estimating equation and quadratic inference function methods for children's respiratory disease and mothers' smoking habits. In each position the first entry is the parameter estimate, the entry in brackets is the estimated standard error, and the third entry is the t-ratio*

| Parameter | Indep(equi) | | | AR(1)(GEE) | | | AR(1)(QIF) | | |
|---|---|---|---|---|---|---|---|---|---|
| Intercept | −1·892 | (0·119) | −15·90 | −1·898 | (0·120) | −15·83 | −1·896 | (0·124) | −15·32 |
| Smoke | 0·305 | (0·188) | 1·62 | 0·275 | (0·190) | 1·45 | 0·280 | (0·185) | 1·51 |
| Age | −0·127 | (0·057) | −2·22 | −0·127 | (0·058) | −2·20 | −0·128 | (0·058) | −2·21 |
| Smoke*Age | 0·056 | (0·088) | 0·64 | 0·062 | (0·089) | 0·69 | 0·048 | (0·087) | 0·54 |

Indep(equi), independent (equicorrelated) correlation; AR(1)(GEE), generalised estimating equation method assuming AR(1) correlation structure; AR(1)(QIF), quadratic inference function method assuming AR(1) correlation structure.

The estimates of the regression parameters are very similar for the two methods, and the $t$-ratios indicate that the child's age, Age, is a significant factor with a negative sign, which means that older children are less likely to get respiratory disease. Maternal smoking, Smoke, contributes positively to children's respiratory disease, though it is not statistically significant; the smoking effect is somewhat inflated if the independence structure is assumed, with $t$-ratio $= 1{\cdot}62$. The interaction between maternal smoking and children's age is insignificant, which implies that there is no indication that the decline in illness differs according to the mother's smoking habit.

The quadratic inference function approach allows us to go beyond individual $t$-tests or 'robust' $z$-tests, and to do a simultaneous test using the statistic in Theorem 1. Table 3 provides chi-squared tests corresponding to the following null hypotheses:

(a) $H_0$: (Smoke, Age, Smoke∗Age) = 0,
(b) $H_0$: (Age, Smoke∗Age) = 0,
(c) $H_0$: (Smoke, Smoke∗Age) = 0,
(d) $H_0$: Smoke∗Age = 0.

In each case, the alternative for the test will be the full model, and the working correlation structure is taken to be AR(1). In Table 3, min $Q$ stands for the minimum of $Q_N(\beta)$ under the null hypotheses, and min $Q_f$ stands for that under the full model. A significant $p$-value indicates that the current model is insufficient to explain the data. It is clear that the most parsimonious model here is just to have the age factor, and again shows that the maternal smoking habit and the interaction term are not significant factors. Moreover, the goodness-of-fit test, with $p$-value 0·331, by Theorem 2 indicates that the model's zero-mean assumption, that is $E(g_N) = 0$, is not rejected.

Table 3. *Children's respiratory disease example. Model selection based on $\chi^2$ test*

| Covariates | min $Q$ | min $Q$ − min $Q_f$ | df | $p$-value |
|---|---|---|---|---|
| Intercept | 13·247 | 8·652 | 3 | 0·034 |
| Intercept, Smoke | 11·192 | 6·597 | 2 | 0·037 |
| Intercept, Age | 6·791 | 2·196 | 2 | 0·334 |
| Intercept, Smoke, Age | 4·903 | 0·308 | 1 | 0·579 |
| Intercept, Smoke, Age, Interaction | 4·595 | 0 | — | — |
| Goodness of fit | 4·595 | — | 4 | 0·331 |

min $Q$, minimum of quadratic inference defined in (6); min $Q_f$, minimum of quadratic inference function for the full model; df, degrees of freedom.

For the following reasons we doubt that the dataset is rich enough to conclude that the effect of maternal smoking habit is not statistically significant: smoking status is treated as fixed rather than as time-dependent, there is no information on the level of maternal smoking habit, and there is no information as to whether or not the mother smokes in the presence of her children. These would be important factors in pursuing more definitive scientific conclusions.

## 5. Discussion

The use of the extended score quadratic inference function approach can be limited by the need for a high-dimensional extended score vector, note that the dimension of $g_N$ is $mq$ instead of $q$, and the specification of a linear approximate inverse. To address these problems, we have developed an adaptive quadratic inference function method, which requires no working assumption; see A. Qu's unpublished Pennsylvania State University Ph.D. dissertation. This approach reduces the dimension of the extended score vector by adding just one moment condition to the quasi-score under independence. The moment conditions are added based on the criterion of increasing the information in the set of estimating functions.

## Acknowledgement

We are grateful for constructive suggestions from an associate editor, the editor and a referee. This research is supported in parts by grants from the National Science Foundation.

## Appendix

### *Proof of Theorem* 1

By Taylor's expansion,

$$Q(\psi_0, \lambda_0) - Q(\hat\psi, \hat\lambda) = \begin{pmatrix} \psi_0 - \hat\psi \\ \lambda_0 - \hat\lambda \end{pmatrix}' \dot Q(\hat\psi, \hat\lambda) + \frac{1}{2} \begin{pmatrix} \psi_0 - \hat\psi \\ \lambda_0 - \hat\lambda \end{pmatrix}' \ddot Q(\psi^\dagger, \lambda^\dagger) \begin{pmatrix} \psi_0 - \hat\psi \\ \lambda_0 - \hat\lambda \end{pmatrix},$$

for some $(\psi^\dagger, \lambda^\dagger)$ between $(\psi_0, \lambda_0)$ and $(\hat\psi, \hat\lambda)$, and

$$Q(\psi_0, \lambda_0) - Q(\psi_0, \tilde\lambda) = (\lambda_0 - \tilde\lambda)' \dot Q_\lambda(\psi_0, \tilde\lambda) + \tfrac{1}{2}(\lambda_0 - \tilde\lambda)' \ddot Q_{\lambda\lambda}(\psi_0, \lambda^*)(\lambda_0 - \tilde\lambda),$$

for some $\lambda^*$ between $\lambda_0$ and $\tilde\lambda$.

Note that $\dot Q(\hat\psi, \hat\lambda)$ and $\dot Q_\lambda(\psi_0, \tilde\lambda)$ are equal to 0 because $(\hat\psi, \hat\lambda)$ and $\tilde\lambda$ satisfy (12). Hence

$$Q(\psi_0, \tilde\lambda) - Q(\hat\psi, \hat\lambda) = \frac{1}{2} \begin{pmatrix} \hat\psi - \psi_0 \\ \hat\lambda - \lambda_0 \end{pmatrix}' \ddot Q(\psi^\dagger, \lambda^\dagger) \begin{pmatrix} \hat\psi - \psi_0 \\ \hat\lambda - \lambda_0 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 0 \\ \tilde\lambda - \lambda_0 \end{pmatrix}' \ddot Q(\psi_0, \lambda^*) \begin{pmatrix} 0 \\ \tilde\lambda - \lambda_0 \end{pmatrix}.$$

We establish the relationship between $(\tilde\lambda - \lambda_0)$ and $(\hat\psi - \psi_0, \hat\lambda - \lambda_0)$ as follows. Expand $\dot Q_\lambda(\psi_0, \tilde\lambda)$ about $\lambda = \lambda_0$ and $\dot Q_\lambda(\hat\psi, \hat\lambda)$ about $(\psi, \lambda) = (\psi_0, \lambda_0)$ to obtain

$$0 = \dot Q_\lambda(\psi_0, \hat\lambda) = \dot Q_\lambda(\psi_0, \lambda_0) + \ddot Q_{\lambda\lambda}(\tilde\lambda - \lambda_0) + O_p(N^{-\frac{1}{2}}),$$

$$0 = \dot Q_\lambda(\hat\psi, \hat\lambda) = \dot Q_\lambda(\psi_0, \lambda_0) + \ddot Q_{\lambda\psi}(\hat\psi - \psi_0) + \ddot Q_{\lambda\lambda}(\hat\lambda - \lambda_0) + O_p(N^{-\frac{1}{2}}).$$

Solving for $\tilde\lambda - \lambda_0$ from the two equations gives

$$(\tilde\lambda - \lambda_0) = \ddot Q_{\lambda\lambda}^{-1} \ddot Q_{\lambda\psi}(\hat\psi - \psi) + (\hat\lambda - \lambda_0) + O_p(N^{-\frac{1}{2}}),$$

where $\dot Q_\lambda = \dot Q_\lambda(\psi_0, \lambda_0)$, $\ddot Q_{\lambda\psi} = \ddot Q_{\lambda\psi}(\psi_0, \lambda_0)$ and so on. That is

$$\begin{pmatrix} 0 \\ \tilde\lambda - \lambda_0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \ddot Q_{\lambda\lambda}^{-1} \ddot Q_{\lambda\psi} & I \end{pmatrix} \begin{pmatrix} \hat\psi - \psi_0 \\ \hat\lambda - \lambda_0 \end{pmatrix}.$$

Then $Q(\psi_0, \tilde\lambda) - Q(\hat\psi, \hat\lambda)$ is asymptotically equivalent to

$$\begin{pmatrix} \hat\psi - \psi_0 \\ \hat\lambda - \lambda_0 \end{pmatrix}^{1} \left\{ \begin{pmatrix} J_{\psi\psi} & J_{\psi\lambda} \\ J_{\lambda\psi} & J_{\lambda\lambda} \end{pmatrix} - \begin{pmatrix} 0 & J_{\psi\lambda} J_{\lambda\lambda}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} J_{\psi\psi} & J_{\psi\lambda} \\ J_{\lambda\psi} & J_{\lambda\lambda} \end{pmatrix} \begin{pmatrix} 0 & 0 \\ J_{\lambda\lambda}^{-1} J_{\lambda\psi} & I \end{pmatrix} \right\} \begin{pmatrix} \hat\psi - \psi_0 \\ \hat\lambda - \lambda_0 \end{pmatrix}$$

$$= (\hat\psi - \psi_0)'(J_{\psi\psi} - J_{\psi\lambda} J_{\lambda\lambda}^{-1} J_{\lambda\psi})(\hat\psi - \psi_0). \quad \text{(A1)}$$

By Theorem 3.2 of Hansen (1982),

$$\begin{pmatrix} \hat\psi - \psi_0 \\ \hat\lambda - \lambda_0 \end{pmatrix} \to N_q \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} J_{\psi\psi} & J_{\psi\lambda} \\ J_{\lambda\psi} & J_{\lambda\lambda} \end{pmatrix}^{-1} \right\},$$

in distribution. Therefore, using the formula for a block matrix inverse, we have

$$(\hat\psi - \psi_0) \to N_p(0, (J_{\psi\psi} - J_{\psi\lambda} J_{\lambda\lambda}^{-1} J_{\lambda\psi})^{-1}),$$

in distribution. By this and (A1) we see that $Q(\psi_0, \tilde\lambda) - Q(\hat\psi, \hat\lambda)$ follows $\chi_p^2$ asymptotically.

The local alternative distribution can be derived by LeCam's Third Lemma (Hall & Mathiason, 1990).

## References

Anderson, T. W. (1969). Statistical inferences for covariance matrices with linear structure. In *Multivariate Analysis II*, Ed. P. Krishnaiah, pp. 55–66. New York: Academic Press.

Anderson, T. W. (1970). Estimation of covariance matrices which are linear combinations or whose inverse are linear combinations of given matrices. In *Essays in Probability and Statistics*, Ed. R. C. Bose et al., pp. 1–24. Chapel Hill, NC: University of North Carolina Press.

Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann Statist.* **1**, 135–41.

Crowder, M. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* **82**, 407–10.

Ferguson, T. S. (1958). A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities. *Ann. Math. Statist.* **29**, 1046–62.

Gine, G. F. & Mason, D. (1997). When is the student $t$-statistic asymptotically standard normal? *Ann. Prob.* **25**, 1514–31.

Hall, W. J. & Mathiason, D. J. (1990). On large-sample estimation and testing in parametric models asymptotic theory. *Int. Statist. Rev.* **58**, 77–97.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–54.

Lee, M. J. (1996). *Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models*. New York: Springer-Verlag.

Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 12–22.

Lindsay, B. G. (1982). Conditional score functions: some optimality results. *Biometrika* **69**, 503–12.

Lindsay, B. G. (1985). Using empirical partial Bayes inference for increased efficiency. *Ann. Statist.* **13**, 914–31.

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

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. New York: Chapman and Hall.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc.* A **135**, 370–84.

Neyman, J. (1949). Contribution to the theory of the $\chi^2$ test. In *Proc. Berkeley Symp. Math. Statist. Prob.*, Ed. J. Neyman, pp. 239–73. Berkeley, CA: University of California Press.

Prentice, R. L. & Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47**, 825–39.

Rao, C. R. (1947). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc. Camb. Phil. Soc.* **44**, 50–7.

Rotnitzky, A. & Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77**, 485–97.

Seely, J. (1971). Quadratic subspaces and completeness. *Ann. Math. Statist.* **42**, 710–21.

Small, C. G. & McLeish, D. L. (1994). *Hilbert Space Methods in Probability and Statistical Inference*. New York: Wiley.

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

[*Received March* 1999. *Revised May* 2000]