Biometrical Journal **40** (1998) 2, 115–139

The Generalised Estimating Equations: An Annotated Bibliography

ANDREAS ZIEGLER

Medical Centre for Methodology and Health Research
Institute of Medical Biometry and Epidemiology
Marburg
Germany

CHRISTIAN KASTNER

Institute of Statistics
LMU München
München
Germany

MARIA BLETTNER

International Agency for Research on Cancer
Lyon Cedex 08
France

Summary

The Generalised Estimating Equations (GEE) proposed by LIANG and ZEGER (1986) and ZEGER and LIANG (1986) have found considerable attention in the last ten years and several extensions have been proposed. In this annotated bibliography we describe the development of the GEE and its extensions during the last decade. Additionally, we discuss advantages and disadvantages of the different parametrisations that have been proposed in the literature. Furthermore, we review regression diagnostic techniques and approaches for dealing with missing data. We give an insight to the different fields of application in biometry. We also describe the software available for the GEE.

Key words: Correlated data analysis; Generalised linear model; Longitudinal data analysis; Marginal model; Pseudo maximum likelihood

Zusammenfassung

Die Generalised Estimating Equations (GEE), die zuerst von LIANG und ZEGER (1986) und ZEGER und LIANG (1986) vorgeschlagen wurden, haben in den vergangenen zehn Jahren große Beachtung gefunden. Verschiedene Erweiterungen wurden vorgeschlagen. In dieser kommentierten Bibliographie beschreiben wir die Entwicklung der GEE und ihrer Erweiterungen während der letzten zehn Jahre. Dar-

über hinaus diskutieren wir Vor- und Nachteile verschiedener in der Literatur vorgeschlagener Parametrisierungen. Wir stellen ebenfalls Ansätze zur Regressionsdiagnostik für GEE sowie zur Behandlung fehlender Daten dar und geben einen Einblick in die Anwendungsgebiete der GEE. Schließlich weisen wir auf Software zur Analyse der GEE hin.

Résumé

Les «Generalised Estimating Equations» (GEE) proposées par LIANG et ZEGER (1986) et ZEGER et LIANG (1986) ont reçu une grande attention pendant les dix dernières années. Cette bibliographie commentée décrit le développement précis des GEE et des extensions qui ont été proposées, discute les variées paramétrisations présentées dans la littérature, et recense les techniques de diagnostic de régression ainsi que les procédés de traitement des valeurs manquates. En outre, nous mettrons au courant des différents domaines d'application de biométrie. Enfin, les logiciels disponibles sont présentés.

1. Introduction

In many biometrical, epidemiological, social and economical situations, the classical assumptions of statistics, in particular the independence of variables and their normal distribution are not valid. For example, count data (like number of epileptic seizures) or binary data (like person being ill or not) are not normally distributed. The independence of outcome variables, for example, is not given when different measurements are taken from the same patient, when he receives several treatments, or when the treatment consists of a number of cycles. The assumption of independence is also violated, if paired data are collected, like for paired organs. Neglecting dependencies in these situations can lead to false conclusions. The precision of the results and thereby their significance is usually overestimated. This is illustrated nicely by SHERMAN and LE CESSIE (1997).

For these reasons, models for the analysis of non-metric correlated data were developed very early. Nevertheless, these models and the technical possibilities of evaluating these models were limited, so that an adequate analysis of relevant questions was not always possible. The development of computer intensive statistical methods like the Generalised Linear Models (GLM; MCCULLAGH and NELDER, 1989) or the Generalised Estimating Equations (GEE; LIANG and ZEGER, 1986) presented here only became possible with the availability of powerful computers. This is also the reason for the increasing interest in the analysis of correlated observations.

Clusters are a way to represent correlated observations: one assumes the existence of a relation between the observations of a cluster, while there is none between observations of separate clusters. Such structures can be induced by the design of the data. Examples are:

- longitudinal or panel data
- family studies
- studies with spatial structures

The clusters themselves do not need to be homogenous: they may have subclusters, as in family studies, where cluster structures emerge from the relationship between parents, between parents and children, and between the children.

The primary interest in the cited examples lies in finding the influence of the variables on a certain key value, the response. Many papers have investigated the situation where the response is continuous and approximately normal. However, the case of binary or categorical dependent variables was only addressed in the last years. The GLM, a generalisation of the regression model for continuous and discrete response, is the classical starting point for current research.

Marginal models, conditional models and random effects models are extensions of the GLM for correlated data (DIGGLE, LIANG and ZEGER, 1994; FAHRMEIR and TUTZ, 1994; KENWARD and JONES, 1992; LIANG and ZEGER, 1986; NEUHAUS, KALBFLEISCH and HAUCK, 1991; ZEGER and LIANG, 1986). A survey of methods for analyzing correlated binary response data (PENDERGAST et al., 1996) as well as a comparison of different approaches for paired binary data (GLYNN and ROSNER, 1994) have been published recently. An annotated bibliography of methods for analyzing correlated categorical data has been given by ASHBY et al. (1992). The GEE (LIANG and ZEGER, 1986; ZEGER and LIANG, 1986) belong to the class of marginal models to which we restrict our attention in this annotated bibliography. We shall give a description how the GEE were developed in the last decade. References point to both the biometrical literature and the econometric literature. Extensions and a caveat are also discussed. Different overviews, in general more theoretical as in this paper, have been given (DAVIS, 1991; FITZMAURICE, LAIRD and ROTNITZKY, 1993; LIANG, 1992; LIANG and ZEGER, 1992; ZEGER, 1988; ZEGER and LIANG, 1992).

Section 2 of this paper cites some examples from the literature to clarify the problem and to show different situations for the application of the GEE. The examples differ in the nature of the dependent variables and the cluster structures.

Section 3 introduces a formal description of cluster structures. The main interest is modelling the expected value of the dependent variables as a function of independent variables. The GLM and the linear model with estimated covariance matrix (feasible generalised least squares (FGLS), feasible Aitken estimator; GREENE, 1993) are well-known examples. The GLM, however, cannot take into account dependencies within clusters, while the linear model requires functional independence of the mean and the variance. The GEE proposed by LIANG and ZEGER (1986) are a synthesis of these models. They generally give asymptotic results. Thus, they require a sufficiently large number of clusters.

Section 4 introduces GEE approaches for situations, where correlations within clusters are to be analysed in addition to the mean structure. In section 5 several extensions of the GEE are briefly introduced and discussed. In section 6 we consider the efficiency of the methods, the bias and the problem of convergence in practical situations. In section 7 an insight to the different fields of application is given. Furthermore, in section 8 we describe the existing software that we are aware of. Finally, we comment on the practical use of the GEE for data analysis.

2. Examples

Example 1: Longitudinal data

The first data set that has been analysed using GEE methods, investigated the stress of mothers in the presence of a child's disease (LIANG and ZEGER, 1986; ZEGER, LIANG and SELF, 1985). The study included 167 mothers with children aged between 18 months and 5 years. Mothers were asked on 28 successive days whether they felt stressed or not. Additionally, an interview was conducted at the beginning of the 28 days period in which some additional questions were asked, concerning health status of the child, marital status, ethnic group and whether the mother was employed or not. The main interest of the investigation was whether these variables had a significant impact on the health status of the child. The marginal mean – a probability – was modelled via the logit model. The binary response variable was whether the child was diseased on day t or not. The correlation due to the repeated measurements from the same child was only of secondary interest.

Example 2: Clinical trial

In this example the effectiveness of an antiepilepticum was investigated. The data have been analysed several times as examples in the literature (e.g. DIGGLE et al., 1994; ZIEGLER and GRÖMPING, 1998). For each patient, the number of epileptic events within 8 weeks was counted before the controlled trial started. All patients were randomised to two treatment groups: additionally to the standard chemotherapy, patients from the first group were given an antiepileptic drug while the second group received a placebo. Response variable was the number of epileptic seizures in 4 two-weeks intervals following the treatment. Additionally, variables such as age of the patients were used for the analysis. The marginal mean was modelled via the Poisson model. Again, the correlation was of secondary interest.

Example 3: Epidemiological study

In order to determine the relative importance family of genetics and environment on the occurrence of athopic disease, a case-control study with 426 patients with athopic disease and 628 controls was carried out yielding overall to some 5000 family members (DIEPGEN and BLETTNER, 1996). The response variable was a binary variable, namely whether athopic disease was present or not. The aim of the study was to investigate presence of a significant association of the disease between parents and their children. If the parameters of the marginal mean should be interpreted as odds ratios, they stroud be modelled via the logit model. This association can give hints whether the disease has a genetic component.

In the first two examples the mean structure was of primary interest while it was the association between family members in the third example. The strong correlation between the persons has to be taken into account in the first two examples to analyse the mean structure. In the last example, several variables that concern either the families or the persons in the family can be used to separate the association of interest from the influence of other covariates. It should be noted that for the first two examples the correlation among the persons may not be neglected to obtain both correct parameter estimates and correct variance estimates for the mean structure.

3. The Generalised Estimating Equations for Estimation of Mean (GEE1)

Let $y_i = (y_{i1}, \ldots, y_{iT})$ be a vector of responses from n clusters, e.g. families or periods, with T observations for the ith cluster, $i = 1, \ldots, n$. For each y_{it} a vector of covariates x_{it} is available, which possibly contains an intercept. The data can be summarised to the vector y_i and the matrix $X_i = (x'_{i1}, \ldots, x'_{iT})'$. The method can be extended to unequal cluster sizes T_i (cf. ZIEGLER and GRÖMPING, 1998). The pairs (y_i, X_i) are assumed to be independently identically distributed. We will first describe models for the mean structure $E(y_{it} \mid x_{it})$. It is necessary to find a method that can deal with the association between the T observations of cluster i. For independent observations, the GLM allows flexibility in modelling mean and variance structures. In GLM, the mean structure is given by $E(y_{it} \mid x_{it}) = \mu_{it} = g(x'_{it}\beta)$, where g is a non-linear response function and β is the unknown $p \times 1$ parameter vector of interest. g^{-1} is termed link function.

We do not consider conditional models, also termed state dependence models $E(y_{it} \mid X_i) = g(x'_{it}\beta + \sum_{t \neq t'} \gamma_{t'} y_{it'})$, where the tth response may depend on responses within the same cluster. Furthermore, we do not consider random effects models, also named mixed models $E(y_{it} \mid X_i) = g(x'_{it}\beta + z'_{it}\gamma_i)$, where γ_i follows some distribution F. Both conditional models and random effects models have been discussed in detail e. g. by FAHRMEIR and TUTZ (1994).

An important property of the GLM is the functional relation between mean and variance $v_{it} = V(y_{it} \mid x_{it}) = h(\mu_{it})\,\phi$. h and ϕ are called variance function and dispersion parameter, respectively. For the purpose of this paper, we set $\phi = 1$ except for the normal distribution, where we use $\phi = \sigma^2$. If a specific univariate exponential family can be assumed, e.g. a normal, Binomial, Poisson or gamma distribution, the variance function is uniquely determined by this assumption. For example, the variance function is constant $(h(\mu_{it}) = 1)$ for the normal, Binomial $(h(\mu_{it}) = \mu_{it}(1 - \mu_{it}))$ for the Binomial, identity $(h(\mu_{it}) = \mu_{it})$ for the Poisson, and squared $(h(\mu_{it}) = \mu_{it}^2)$ for the gamma distribution. Further examples of link and variance functions are given e.g. by MCCULLAGH and NELDER (1989).

For independent observations, the parameter vector β is estimated using the maximum likelihood (ML) method: The distribution – e.g. the Binomial or Poisson distribution – determines the likelihood equations (score equations) that are

given by derivatives of the log-likelihood function with respect to β. The score equations have the form

$$u(\beta) = \frac{1}{n} \sum_{i=1}^{n} D_i' \Sigma_i^{-1} (y_i - \mu_i) = \tfrac{1}{n} D' \Sigma^{-1} (y - \mu) = 0\,, \tag{1}$$

where $D_i = \partial \mu_i / \partial \beta'$ is the diagonal matrix of first derivatives and Σ_i is the diagonal matrix of the variances $\Sigma_i = \text{diag}\,(v_{it})$. Furthermore, D and y are the stacked D_i matrices and y_i vectors, respectively. Σ is the block diagonal matrix of the Σ_i, μ_i is the vector of μ_{it}, and μ is defined analogously to y. (1) are called independence estimating equations (IEE; LIANG and ZEGER, 1986). In general, (1) has to be solved iteratively by Fisher's scoring algorithm, iterative weighted least squares (IWLS) or Quasi-Newton algorithms (LUENBERGER, 1984). The estimator $\hat{\beta}$ is consistent and asymptotically normal distributed with covariance matrix $V(\hat{\beta}) = (D' \Sigma^{-1} D)^{-1}$.

For correlated observations, however, the true variance matrix is not diagonal. If the conditional variance matrix Σ_i does not equal the true variance matrix $V(y_i \,|\, X_i) = \Omega_i$, the estimator $\hat{\beta}$ still remains unbiased. For consistent estimation of $V(\hat{\beta})$ the robust variance matrix, also termed sandwich information matrix, should be used instead of the Fisher information matrix. The robust variance matrix estimator traces back to HUBER (1967) and has been further examined (GOURIEROUX, MONFORT and TROGNON, 1984; LIANG and ZEGER, 1986; ROYALL, 1986; WHITE, 1982; ZEGER and LIANG, 1986; ZEGER et al., 1985):

$$\begin{aligned} \hat{V}(\hat{\beta}) = \hat{H}_1^{-1} \hat{H}_2 \hat{H}_1^{-1} = &\left(\sum_{i=1}^{n} \hat{D}_i' \hat{\Sigma}_i^{-1} \hat{D}_i \right)^{-1} \left(\sum_{i=1}^{n} \hat{D}_i' \hat{\Sigma}_i^{-1} \hat{\Omega}_i \hat{\Sigma}_i^{-1} \hat{D}_i \right) \\ &\times \left(\sum_{i=1}^{n} \hat{D}_i' \hat{\Sigma}_i^{-1} \hat{D}_i \right)^{-1}, \end{aligned} \tag{2}$$

where $\hat{\Omega}_i = (y_i - \hat{\mu}_i)\,(y_i - \hat{\mu}_i)'$. Note that $\hat{\Omega}_i$ is not a suitable estimator of Ω_i. $\hat{\mu}_i$ is defined by the link function of the GLM. $\hat{H}_1$ is the estimated Fisher information matrix. $\hat{H}_2$ is termed estimated outer product gradient (OPG) because it is the estimate of the expected outer product of the score vector. BINDER (1983) proposed an estimator similar to (2) based on a taylor linearisation for implicitly-defined parameters so that the outer matrices are not necessarily symmetric.

In several situations the estimation will not be very efficient because of the diagonal form of Σ_i. The GEE1 approach allows more efficient estimation: Consider a GLM with fixed mean structure and variance. In this case Σ_i is a covariance matrix which should be close to the true covariance matrix Ω_i. Keep in mind that the association (correlation) is not of primary interest here. The $T \times T$ correlation matrix $R(\alpha)$ of y_i given X_i, well-described by an additional parameter (vector) α, is assumed to be identical for all clusters. For example, one has $\text{Corr}\,(y_{it}, y_{it'} \,|\, X_i) = \alpha$ for $t \neq t'$ in an equicorrelated model. If $R(\hat{\alpha}) = \hat{R}$ is an esti-

mator of the correlation matrix, the estimator for Σ_i is given by

$$\hat{\Sigma}_i = \hat{A}_i^{1/2} R(\hat{\alpha})\, \hat{A}_i^{1/2}\,, \tag{3}$$

where $\hat{A}_i^{1/2}$ is the estimated square root of the diagonal matrix of the variances v_{it}.

With the estimated working correlation matrix $\hat{R}$ and the diagonal matrices $\hat{A}_i$, the GEE1 have the form

$$u(\beta) = \frac{1}{n} \sum_{i=1}^{n} D_i' \hat{\Sigma}_i^{-1} (y_i - \mu_i) = 0\,. \tag{4}$$

The term 'generalised' is somehow misleading. However, it is justified considering that LIANG and ZEGER (1986) developed the equation system (4) from the GLM or the IEE, respectively. Nowadays, the term Estimating Equations (EE) is preferred to GEE (PRUSCHA, 1996). GEE1 means that only first order moments, i.e. the mean structure, are estimated consistently.

LIANG and ZEGER (1986) and ZEGER and LIANG (1986) used the method of moments to estimate the 'working correlation matrix' $R(\alpha)$. The choice of $R(\alpha)$ was discussed for example by LIANG and ZEGER (1986) in some detail (s. also ZIEGLER and GRÖMPING, 1998). If the identity matrix is used as working correlation matrix, (3) is reduced to a diagonal matrix for the variances, and the EE (4) reduce to the IEE (1). The working correlation matrix needs to be chosen carefully. If it is not well-specified, existence of $\hat{\alpha}$ and convergence of $\hat{\alpha}$ to α cannot be ensured (CROWDER, 1995). Furthermore, even if $\hat{\alpha}$ converges to some fixed value α assuming an equicorrelated structure, interpretation might be difficult or even impossible, if the true correlation structure is autoregressive. Note that (4) is similar to the Feasible Generalised Least Squares (FGLS) estimator (COCHRANE and ORCUTT, 1949; GREENE, 1993) where in the first step the variance matrix $\hat{\Sigma}_i$ and in the second step the parameter vector β are estimated. Usually, the GEE1 are solved by a modified Fisher scoring algorithm. The term 'modified' indicates that $\hat{\Sigma}_i$ is used instead of the true covariance matrix for solving (4). In order to save CPU-time, LIPSITZ et al. (1994b) proposed to use a one-step approximation of (4) with the assumption of independence as starting value.

If β is estimated using (4), $\hat{\beta}$ is consistent under suitable regularity conditions, if $\mu_{it} = E(y_{it} \,|\, x_{it}) = E(y_{it} \,|\, X_i)$ is specified correctly. The GEE1 estimator is asymptotically normal. The variance can be estimated consistently with the robust variance estimator (2) and $\hat{\Sigma}_i$ as in (3). The required regularity conditions can be formulated either by embedding the GEE into the framework of Quasi Likelihood estimation (QL; FIRTH, 1993; MCCULLAGH and NELDER, 1989), as shown by ROTNITZKY (1988), or by embedding the GEE into the framework of Pseudo Maximum Likelihood estimation (PML; GOURIEROUX et al., 1984; GOURIEROUX and MONFORT, 1993) or by embedding the GEE into the Generalised Method of Moments (GMM; HANSEN, 1982; NEWEY, 1993) as shown by ZIEGLER (1995). If Σ_i is specified correctly, i.e. $\Omega_i = \Sigma_i$, and the true distribution belongs to the linear exponential family, $\hat{\beta}$ is efficient in the sense of Rao-Cramér (GOURIEROUX and MONFORT, 1993).

Most estimators for the correlation structure R can be developed using EE (CROWDER, 1995; ZIEGLER, 1994). It follows that additionally to the EE for β a second set of EE for α can be introduced. The general form of this EE system is (PRENTICE, 1988)

$$u(\alpha) = \frac{1}{n} \sum_{i=1}^{n} E_i' \Psi_i^{-1} (z_i - \varrho_i(\alpha)) = 0\,. \tag{5}$$

In (4) the expectation μ_i of y_i, is given as a function of the parameter β. In (5) the vector form $\varrho_i(\alpha)$ of the correlation matrix $R(\alpha)$ is given as a function of the parameters of association α. $z_{itt'}$ is the product of the Pearson residuals and includes observations as well as parameters: $z_{itt'} = (y_{it} - \mu_{it})\,(y_{it'} - \mu_{it'})/\sqrt{v_{it} v_{it'}}$. z_i is of dimension $T(T-1)/2$ and is defined analogously to the response vector y_i. E_i is the matrix including the first derivatives of $\varrho_i(\alpha)$ with respect to α. Ψ_i^{-1} can be interpreted as the inverse of the covariance matrix of z_i.

The advantage of using (5) compared to the use of the method of moments is that non-linear correlation structures can be estimated. Similar to the link function in GLM, we can define the association with explanatory variables X_i in the form $\varrho_i(\alpha) = \varrho_i(X_i, \alpha)$ (LIPSITZ, LAIRD and HARRINGTON, 1991). However, it is not straight forward to define a reasonable function to model the association between the correlation structure $\varrho_i(X_i, \alpha)$ and the covariates X_i (LIPSITZ et al., 1991; LIPSITZ et al., 1994b). The problem is that the covariates can be continuous but the correlations are restricted to the interval $[-1; 1]$. Therefore it is commendable to define restrictions for the correlation structure which should be non-linear functions in analogy to the well-known link function. An example for such a function is the area tangens hyperbolicus. The transformation has a similar interpretation as the link function in GLM and we thus call it association link function.

4. Generalised Estimating Equations for Estimation of Mean and Association (GEE2)

In the last section we considered EE that allow consistent estimation of the mean. We will now describe a set of EE that permit consistent estimation of the parameters of first and second order moments. These EE are called GEE2. Currently, no clear and unique definition of GEE2 is possible, as several procedures are summarised by this term. LIANG, ZEGER and QAQISH (1992) used the phrase GEE2 for simultaneous estimation of the mean and the association. We shall name them EE of first and second order. In our terms, first and second order EE might be solved separately.

4.1 *The GEE using the correlation as the measure of association*

The two systems (4) and (5) have a comparable form. They can be summarised to:

$$u\begin{pmatrix}\beta\\ \alpha\end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} \dfrac{\partial \mu_i}{\partial \beta'} & 0 \\ 0 & \dfrac{\partial \varrho_i}{\partial \alpha'} \end{pmatrix}' \begin{pmatrix} V(y_i) & 0 \\ 0 & V(z_i) \end{pmatrix}^{-1} \begin{pmatrix} y_i - \mu_i \\ z_i - \varrho_i \end{pmatrix} = 0\,. \tag{6}$$

It can be seen that the matrix of first derivatives and the working covariance matrix are block-diagonal. Therefore, (6) is a simplification of the following system:

$$u\begin{pmatrix}\beta\\ \alpha\end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\frac{\partial\mu_i}{\partial\beta'} & \frac{\partial\mu_i}{\partial\alpha'}\\ \frac{\partial\varrho_i}{\partial\beta'} & \frac{\partial\varrho_i}{\partial\alpha'}\end{pmatrix}'\begin{pmatrix}V(y_i) & \mathrm{Cov}\,(y_i, z_i)\\ \mathrm{Cov}\,(z_i, y_i) & V(z_i)\end{pmatrix}^{-1}\begin{pmatrix}y_i-\mu_i\\ z_i-\varrho_i\end{pmatrix} = 0\,. \qquad (7)$$

The form $\partial\mu_i/\partial\alpha' \neq 0$ in (7) implies that the association – here the correlation – is a function of β. This assumption is not plausible because it is difficult to interpret a mean vector that includes the association parameter α. In applications, the mean values are only defined as a function of β, which implies $\partial\mu_i/\partial\alpha' = 0$. In practice, ϱ_i is usually defined via the area tangens hyperbolicus. Thus, ϱ_i is independent of β so that $\partial\varrho_i/\partial\beta' = 0$. If the matrix of first derivatives is block-diagonal, the covariance matrix of (7) has to be block diagonal, to guarantee consistent estimators $\hat{\beta}$ of β (PRENTICE and ZHAO, 1991). Hence, (7) reduces to (6), if ϱ_i is modelled via the area tangens hyperbolicus.

The EE (6) and (7) can be embedded into the GMM (ZIEGLER, 1995) so that the estimators $\hat{\beta}$ and $\hat{\alpha}$ are jointly asymptotically normal under regularity conditions formulated by HANSEN (1982). The asymptotic covariance matrix corresponding to (6) is given by PRENTICE (1988). The EE (6) and (7) may be solved by a modified Fisher scoring algorithm analogously to the GEE1.

4.2 *The GEE using the covariance as the measure of association*

So far, we only defined EE using the correlation as the measure of association. The EE can also be defined using the covariance matrix. Then $s_{itt'} = (y_{it} - \mu_{it}) \times (y_{it'} - \mu_{it'})$ and $\sigma_{itt'} = E(s_{itt'}) = \mathrm{Cov}\,(y_{it}, y_{it'})$ are used instead of $z_{itt'}$ and $\varrho_{itt'}$, respectively, in (7). The first derivatives and the working variance matrices are changed accordingly. This approach was first proposed by ZHAO and PRENTICE (1990), and is closely related to the method described by CROWDER (1985). The main question is, how to model the association between s_i and α and s_i and β, respectively. $\sigma_{itt'}$ can be modelled as a function of β via v_{it} and as a function of α via $\varrho_{itt'}$ since $\sigma_{itt'} = (v_{it}v_{it'})^{-1/2}\,\varrho_{itt'}$.

If μ_i and σ_i are correctly specified as functions of α and β, the estimates $\hat{\alpha}$ and $\hat{\beta}$ are consistent and jointly asymptotically normal (GOURIEROUX and MONFORT, 1993; PRENTICE and ZHAO, 1991; ZHAO and PRENTICE, 1990, 1991) with asymptotic covariance matrix given e.g. by PRENTICE and ZHAO (1991). These EE are not often applied, due to the following disadvantage compared to (6): It is necessary to specify μ_i as well as σ_i correctly to obtain a consistent estimate $\hat{\beta}$. If the correlation is used instead of the covariance, the estimator $\hat{\beta}$ remains consistent, even if $\varrho_i(\alpha)$ is incorrectly specified via the arcus tangens hyperbolicus because $\hat{\alpha}$

and $\hat{\beta}$ are estimated in separate EE. The advantage of this two-step procedure was first observed by FIRTH (1992) and DIGGLE (1992) in their discussions of the paper by LIANG et al. (1992).

4.3 *The GEE using the second ordinary moments as the measure of association*

We shall now sketch an approach that is only applicable to dichotomous or categorical response variables. In this situation, the relationship between the second moments and the log odds ratios is well-known (BISHOP, FIENBERG and HOLLAND, 1975) and the log odds ratio can be modelled as linear functions of the covariates X_i and the unknown parameter α.

If (7) is used in the log odds ratio parameterisation, consistent estimators $\hat{\beta}$ and $\hat{\alpha}$ exist and are jointly asymptotic normal, if both the mean and the association structure are specified correctly (PRENTICE and ZHAO, 1991; ZHAO and PRENTICE, 1990, 1991). The asymptotic covariance matrix is given in LIANG et al. (1992). Misspecification of α can lead to an inconsistent estimate of β, since $\hat{\beta}$ and $\hat{\alpha}$ are estimated simultaneously.

The simultaneous estimation procedure for α and β can be transformed into a two-step procedure. Then $\hat{\beta}$ is consistent, even if α is incorrectly specified. This approach is called 'alternating logistic regression' (ALR; CAREY, ZEGER and DIGGLE, 1993) with the logit-link as link function. The ALR is closely related to the approach of PRENTICE (1988).

4.4 *The GEE using the polychoric and polyserial correlation as the measure of association*

The GEE approach using polychoric and polyserial correlations as the measure of association can be derived using latent variable models which are commonly used in econometrics, while the GEE have mostly been applied in biometry. It was considered in detail e.g. by LE CESSIE and VAN HOUWELINGEN (1994), QU et al. (1992) or ZIEGLER and ARMINGER (1995). It is closely related to the Mean and Covariance Structure analysis (BROWNE and ARMINGER, 1995). Thus it might be applied to mixtures of continuous, dichotomous, categorical and limited dependent variables. If the marginal probabilities are not too close to 1 or 0, the approach using the polychoric correlation should yield similar results as the log odds ratio approach discussed in section 4.3 (LE CESSIE and VAN HOUWELINGEN, 1994).

5. Extensions of the Generalised Estimating Equations

5.1 *Time dependent parameters*

One limitation of the GEE approaches is that the parameter vector β has to be constant for all t. GEE can be extended to include a time dependent parameter

vector. This extension is important for longitudinal studies where the influence of the covariates changes with time (LIPSITZ, KIM and ZHAO, 1994c; WEI and STRAM, 1988; ZIEGLER and ARMINGER, 1995). The basic idea of this approach is to rearrange the explanatory variables x_{it} in a matrix

$$X_i' = \begin{pmatrix} x_{i1} & 0 & \dots & 0 \\ 0 & x_{i2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & x_{iT} \end{pmatrix}. \tag{8}$$

The parameter vector β is given by $\beta = (\beta_1', \dots, \beta_T')'$. The estimation itself proceeds as above. A similar approach can be used for joint estimation of time varying and time constant parameters (PARK, 1994).

5.2 *Ordered categorical and non-ordered categorical dependent variables*

The GEE can be extended to ordered categorical and non-ordered categorical data. The basic idea is to apply analogies of the multivariate logit (probit) model or the cumulative logit (probit) model, also termed proportional odds model. As in the last section, the explanatory variables need to be rearranged. In addition, the categorical response y_{it} has to be recoded. For example, consider an ordinal response y_{it} with four possible categories. Then three thresholds (cutpoints) are required for the cumulative model which correspond to an intercept and two additional dummy variables. The independent variables are arranged to

$$X_i' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & \dots & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & & 0 & 1 & 0 \\ x_{i1} & x_{i1} & x_{i1} & x_{i2} & x_{i2} & x_{i2} & & x_{iT} & x_{iT} & x_{iT} \end{pmatrix}. \tag{9}$$

Three dummy variables without an intercept could be used instead. The dependent variable y_{it} has to be dummy coded and results to one of the four vectors $(1, 0, 0)'$, $(0, 1, 0)'$, $(0, 0, 1)'$, $(0, 0, 0)'$ depending on the value of y_{it}. The working correlation matrix should take into account the correlation structure of the multinomial distribution. Details can be found e.g. in GANGE et al. (1994), HEAGERTY and ZEGER (1996), KENWARD, LESAFFRE and MOLENBERGHS (1994a, 1994b), LIPSITZ et al. (1994c), LUMLEY (1996a), MILLER (1995), MILLER, DAVIES and LANDIS (1993), O'HARA HINES (1997b, 1998), STRAM, WEI and WARE (1988), WILLIAMSON, KIM and LIPSITZ (1995) and ZIEGLER (1994).

5.3 *Missing data*

In many applications one is concerned with missing data. The methods described above yield only unbiased estimates for the mean structure if the data are missing

completely at random (MCAR; RUBIN, 1976). A general approach for calculating the magnitude of the bias of estimators obtained from standard analysis of EE in the presence of incomplete data was presented by ROTNITZKY and WYPIJ (1994). Several approaches have been proposed to deal with missing data in the framework of the GEE.

The first approach is based on the EM algorithm (DEMPSTER, LAIRD and RUBIN, 1977) and may be applied to the GEE1 approach, if dependent variables y_{it} are missing. The X_i's have to be observed completely. The idea is that the GEE1 can be interpreted as EE of a multivariate normal distribution with mean μ_i and variance Σ_i. Then the EM algorithm for normally distributed data (JENNRICH and SCHLUCHTER, 1986) may be applied (MAY and JOHNSON, 1995), which yields consistent parameter estimates, if the data are missing at random (MAR).

The second approach is based on the framework of PML estimation and has been proposed by ZIEGLER (1994) for the GEE1 as an extension of the work by ARMINGER and SOBEL (1990). It is a computationally simple approach and may be applied in the presence of missing dependent variables, if the data are MCAR. This approach is in general more efficient than the usually applied complete case analysis. The basic idea is to use EE of a density with complete data that ensures positive definiteness of $\hat{\Sigma}_i$, and that is proportional to the density of the incomplete data.

The third approach is an extension of a traditional approach by KOCH, IMREY and REINFURT (1977) to missing data for categorical dependent variables. As before, it is assumed that the X_i's are completely observed. The approach has been proposed by LIPSITZ, LAIRD and HARRINGTON (1994d) and yields consistent estimates, if the data are MAR. It is a two-step method that applies the EM algorithm in the first step to obtain unrestricted estimates of multinomial probabilities. In the second step, β is estimated using the estimated response vectors.

A fourth approach proposed by Robins and co-workers in a series of papers (ROBINS, ROTNITZKY and ZHAO, 1994, 1995; ROBINS and ROTNITZKY, 1995; ROTNITZKY and ROBINS, 1995a, 1995b) has received considerable attention. It may be applied to both the GEE1 and the GEE2 in the presence of missing dependent variables and/or missing independent variables, if the data are MAR. The idea of this approach is to use weighted EE (WEE) similar to the well-known Horvitz-Thompson estimation. ROBINS and ROTNITZKY (1995a) show that the WEE are efficient in the sense of NEWEY (1990). They use a different variance estimator because the robust variance estimator (2) may not be positive definite (ROBINS et al., 1995). ZHAO, LIPSITZ and LEW (1996b) proposed joint estimating equations (JEE) as a special case of the WEE for missing covariates. XIE and PAIK (1997a) extended the WEE for missing covariates to longitudinal data. An extension to non-ignorable missing-data mechanisms has recently been proposed (ROTNITZKY and ROBINS, 1997; TROXEL, LIPSITZ and BRENNAN, 1997).

Finally, XIE and PAIK (1997b) and PAIK (1997) extended the multiple imputation method to the case of longitudinal data.

5.4 *Testing hypothesis and regression diagnostics*

In many applications testing hypothesis about certain parameters is of substantial interest. The classical asymptotic Gauss test with robust standard errors may be applied to evaluate the significance of a single parameter. This approach may also be used to construct confidence intervals. Instead of applying the asymptotic Gauss test, one could use an added variable plot to check wether an omitted variable should be included in the model (HALL, ZEGER and BANDEEN-ROCHE, 1994).

Two different approaches have been proposed for testing complex hypothesis in the mean and/or the association structure. The approach of ROTNITZKY and JEWELL (1990) is directly based on the GEE1 of LIANG and ZEGER (1986). These authors propose a modified Wald statistic, a modified score statistic and a model based likelihood ratio statistic. They derive the asymptotic distributions of these statistics under the null and the alternative hypotheses. The second approach is based on test statistics derived in the framework of PML estimation (ARMINGER, 1992; ZIEGLER, 1994). In this approach a modified score statistic, a modified Wald statistic and some measures for goodness of fit are proposed. The Wald and the score test statistics are similar to those used in the framework of ML estimation. Here, the model based variance matrix is replaced by the robust variance matrix.

Originally, the GEE were derived to avoid complete specification of the likelihood. Thus, it is impossible to correctly specify the likelihood function in many applications. In these situations, the likelihood-ratio statistic is a weighted sum of independent χ^2-variables (LIANG and SELF, 1996). Alternatively, one might apply either the score or the Wald test statistic. Confidence ellipsoids may also be constructed in the usual way on the basis of these statistics.

Diagnostic techniques that are used in the linear model or in GLMs (MCCULLAGH and NELDER, 1989) can be carried over to mean structure models that are estimated by the GEE. However, one has to distinguish between observation and cluster specific diagnostic measures. Ordinary, standardised and studentised residuals can be used to check for systematic variation that are caused by one or more regressors (TAN, QU and KUTNER, 1997; ZIEGLER and ARMINGER, 1996). In addition, an empirical residual for a cluster using the Mahalanobis distance may be defined (TAN et al., 1997). A modified hat matrix and a modified Cook-statistic are proposed in order to find leverage and influential points (PREISSER and QAQISH, 1996; TAN et al., 1997; ZIEGLER and ARMINGER, 1996). The standardised, studentised and empirical residuals as well as the modified hat matrix rely on the correct specification of the association matrix.

A simulated Q-Q plot and a half-normal probability plot may be used to detect outliers and to investigate model adequacy. In addition, partial residual plots may be used to evaluate linearity and to provide guidance on how to improve the goodness of fit of the model (TAN et al., 1997). A stepwise model selection procedure has been proposed by NUAMAH, QU and AMINI (1996).

All diagnostic measures can be obtained by one-step approximations. Their computation is fast. Therefore, regression diagnostics should be routinely applied in data analysis.

5.5 *Further extensions*

In the preceding sections the most important extensions of the GEE were outlined. However, there are a few more extensions which should be noted. Originally, the GEE was derived from Quasi Likelihood estimation that captures an additional dispersion parameter, commonly denoted by ϕ. In their original formulation, ϕ is assumed to be equal for all t. PARK (1993) noted that this assumption does not hold in most longitudinal studies. PARK (1993) and PAIK (1992) extended the approach of LIANG and ZEGER (1986) to allow for varying dispersion parameters ϕ_t. PARK and SHIN (1995) compared the efficiency of the original approach and the approach proposed by PARK (1993).

HALL and SEVERINI (1995) extended the GEE2 approach. They formulated joint EE for β, α and ϕ based on Quasi Likelihood and showed in an example that these EE may be more efficient than the GEE2. Note that the dispersion parameter ϕ was not used in the original formulation of the GEE2. QAQISH and LIANG (1992) extended the GEE2 to allow regression structures that include multiple classes and multiple levels of nesting as they occur e.g. in family studies. All models discussed above only allow the inclusion of one type of dependent variables. An extension to mixtures of continuous and dichotomous dependent variables has been proposed by FITZMAURICE and LAIRD (1995). O'HARA HINES (1997a) recently proposed an approach for the analysis of retrospectively sampled clusters with known sampling rates. An extension of the GEE to estimate quantiles instead of the mean structure was proposed by LIPSITZ et al. (1997). Approaches for sample size calculations were developed by LIU and LIANG (1997) and SHIH (1997).

6. Efficieny Considerations, Bias, Convergence and Limitations

All asymptotic properties of the GEE1 require correct specification of the mean structure. An implicit assumption of the GEE is that the investigator has to specify the mean $E(y_{it} \mid X_i)$ of y_{it} given all explanatory variables of the cluster instead of modelling just $E(y_{it} \mid x_{it})$. In practice, μ_{it} is modelled via x_{it} as a function of β. If this implicit assumption is not valid, results are biased. Hence, one needs to validate $E(y_{it} \mid X_i) = E(y_{it} \mid x_{it})$ or should apply the IEE (PEPE and ANDERSON, 1994). However, the use of the IEE instead of the GEE might lead to a decrease of efficiency in the parameter estimates. The efficiency of the IEE compared to the GEE1 has been examined analytically (FITZMAURICE, 1995; LEE, SCOTT and SOO, 1993a; MANCL and LEROUX, 1996) and by simulations (EMRICH and PIEDMONTE,

1992; GUNSOLLEY, GETCHELL and CHINCHILLI, 1995; LIANG and ZEGER, 1986; MCDONALD, 1993; PAIK, 1988; SHARPLES and BRESLOW, 1992). The simulation studies gave inconsistent results that can be explained by the work of FITZMAURICE (1995) and MANCL and LEROUX (1996). These authors showed that efficiency depends on the covariate distribution, the cluster sizes, the regression parameters and the correlation between the responses. The results are quite sensitive to the between-cluster and the within-cluster correlation of the covariates. They showed that for specific models the IEE are as efficient as the GEE, if responses within clusters are independent, or if all covariates within clusters are constant, or if all covariates are mean-balanced, i.e. the cluster means are constant across clusters.

If the matrix of independent variables is quadratic and regular, then the GEE and the IEE are identical. In this situation, they are efficient (SPIESS and HAMERLE, 1996). However, the IEE can be quite inefficient, if there is some within-cluster covariate variation and some inbalance in covariate patterns across clusters.

The efficiency of GEE2 estimation has not been investigated in detail. Some theoretical results exist for the asymptotic distributions in the context of PML estimation (GOURIEROUX and MONFORT, 1993) and of GMM estimation (NEWEY, 1990). All estimation procedures – including ML – yield an underestimation of the covariance matrix (LEE et al., 1993).

The efficiency of ML compared to GEE is not entirely clear. In general, small sample sizes yield biased estimates. This bias decreases with the number of clusters n (SHARPLES and BRESLOW, 1992). A comparison of the unweighted GEE and the WEE proposed by ROBINS et al. (1995) was given by FITZMAURICE, MOLENBERGHS and LIPSITZ (1995).

One major advantage of the IEE is that the algorithm converges in most applications. If additional parameters are included in the model, like in the GEE or the ML models, algorithms converge less often. In addition, the GEE2 diverges more often than the GEE1. To apply GEE2, a simple structure of the working matrix is recommended. If convergence problems occur, it is recommended to set the third moments to 0. Further simplification is obtained by using the identity matrix as lower right block of the working matrix. Note that the extensions to time dependent parameters and/or categorical data may lead to convergence problems due to an increased matrix X.

Marginal models, in practice mainly used for binary variables, have one major disadvantage which can have an important influence in the analysis of categorical data. The parameter space of the association parameter – defined by the correlation or the log odds ratio – is bounded for $T \geq 2$ (FITZMAURICE and LAIRD, 1993; LIANG et al., 1992; PRENTICE, 1988). Similarly, it can be shown that the parameter space of the odds ratios is restricted for $T \geq 3$ (LIANG et al., 1992). A possible solution to this problem is to investigate the full likelihood as proposed by FITZMAURICE and LAIRD (1993) and FITZMAURICE, LAIRD and LIPSITZ (1994). Their approach can be interpreted as an extension of the partly exponential model discussed by ZHAO, PRENTICE and SELF (1992).

7. Application of the GEE in Biometry

The GEE have been applied in various biometrical fields, e.g. in teratological and toxicity studies (BIELER and WILLIAMS, 1995; BOWMAN, CHEN and GEORGE, 1995; RYAN, 1992; ZHU, KREWSKI and ROSS, 1994), in ophthalmologic trials (FRAMINGHAM, 1996a; GANGE et al., 1994; PODGOR et al., 1996), in diagnostic testing programs (LEISENRING, PEPE and LONGTON, 1997) or in the analysis of bioequivalence (TEN HAVE and CHINCHILLI, 1995). It is beyond the topic of this paper to give a complete review of all fields of application. However, to give an idea on the broad use of the GEE, we focus on the application to family studies.

Several approaches have been proposed to establish familial aggregation of a disease. The question how to deal with different ascertainment schemes has not been solved completely. For case-control studies, LIANG and BEATY (1991), TOSTESON, ROSNER and REDLINE (1991), ZHAO and LE MARCHAND (1992) proposed to exclude the proband from the analysis. Recently, WHITTEMORE (1995) and ZHAO et al. (1996a) discussed models that allow inclusion of probands. LIANG and PULVER (1996) have derived sample size formulas for family studies, if the GEE are applied in an unmatched case-control study. The detection of influential families using GEE has been illustrated by ZIEGLER et al. (1998). The GEE has also been proposed to detect linkage (AMOS, 1994; AMOS, ZHU and BOERWINKLE, 1996; OLSON, 1994a; OLSON, 1995; OLSON and WIJSMAN, 1993; ZIEGLER and KASTNER, 1997), to estimate allele frequencies (OLSON, 1994b; ZIEGLER and KASTNER, 1997) and association parameters (AMOS, 1994; OLSON, 1994b; TRÉGOUËT, DUCIMETIÈRE and TIRET, 1997), anticipation (POLITO et al., 1996; SCHNEIDER et al. 1998), heritability (GROVE, ZHAO and QUIAOIT, 1993) and in segregation analyses (LEE, STRAM and THOMAS, 1993b; LEE and STRAM, 1996; STRAM, LEE and THOMAS, 1993; WHITTEMORE and GONG, 1994; ZHAO, 1994).

8. Software

Several programs are available to apply the GEE. Most of these programs are written in program languages that facilitate matrix languages. A SAS IML macro for analysing GEE1 written by KARIM and ZEGER (1988) and extended by GRÖMPING (1993) is available at statlab.uni-heidelberg.de. Other program using the facilities of SAS have been presented by LIPSITZ and HARRINGTON (1990), NUAMAH et al. (1996) and ANDOH and UWOI (1995). An SPSS macro for solving the GEE1 by DUNCAN et al. (1995) that is based on the original SAS macro by KARIM and ZEGER (1988) is available from ftp.ori.org/pub/terryd. S-Plus functions written by NORLEANS (1995) are available at http://fisher.stat.unipg.it/pub/stat/statlib/S/geex. Similar S-Plus functions for the GEE1 and the ALR written by CAREY (1989) can be obtained from lib.stat.cmu.edu. At this site a GENSTAT program is also available (KENWARD and SMITH, 1995a, 1995b). A PASCAL program for the GEE2 of

LIANG et al. (1992) is available either at statlab or at statlib. A FORTRAN program written by DAVIS (1993) can be obtained from the Department of Preventive Medicine, University of Iowa. A XLISP-Stat tool by LUMLEY (1996b) for the GEE and the regression diagnostics by PREISSER and QAQISH (1996) can be obtained from http://www.biostat.washington.edu/~thomas/gee.html. A DOS/Windows program written by the quantitative genetic epidemiology group of the FHCRC (QGE, 1994) for solving GEE is available at mule.fhcrc.org or statlab.uni-heidelberg.de. MAREG – a DOS/Windows and SunOS program – for solving the GEE1 and the GEE2 using the approach of PRENTICE (1988) and the WEE approach of ROBINS et al. (1995) for monotone missing data patterns is available from http://www.stat.uni-muenchen.de/~andreas/winmareg.html (KASTNER, FIEGER and HEUMANN, 1996).

Recently, the GEE1 were integrated into procedures of the commercially available program systems SAS, release 6.12 (PROC GENMOD), Stata, release 5.0 (procedure XTGEE), SUDAAN, release 7.11 (PROC MULTILOG), and SPIDA, release 6 (procedure GEE). A comparison of SAS (PROC GENMOD), Stata (XTGEE) and SUDAAN (PROC MULTILOG) is given by ZIEGLER and GRÖMPING (1998).

If one cannot facilitate any of these programs, one can approximate the robust variance matrix by using jackknife techniques (LIPSITZ, LAIRD and HARRINGTON, 1990; LIPSITZ, DEAR and ZHAO, 1994a; PAIK, 1988; PREGIBON, 1983; ZIEGLER, 1997) or by a nonparametric bootstrap (SHERMAN and LE CESSIE, 1997). These are appealing approaches to obtain estimates of the robust variance matrix e.g. in survival models.

9. Discussion and Recommendations

For practical use, some recommendations are required to decide whether the ML methods for multivariate distributions (e. g. FITZMAURICE and LAIRD, 1993; FITZMAURICE et al., 1994) or GEE methods should be used. In general, the ML method should only be applied, if the complete distribution of y_i given X_i can be specified correctly. Otherwise, misspecification may yield inconsistent estimates of the parameters. These inconsistencies may affect either only the asymptotic variance matrix or both, the parameters and their asymptotic variance matrix. GEE1 yields consistent estimates, if the mean structure is correctly specified. However, the association between observations within clusters is treated as nuisance parameter. The use of the robust estimators for the variance is recommended, if misspecification of the association structure is possible. If the investigation of the association is the main goal of the analysis, GEE2 can be applied. However, both the mean and the association structure have to be specified correctly in this situation. If block diagonal matrices are used, GEE2 yields consistent estimates of the mean-structure, even if the association is not specified correctly. Note that the ML approach of FITZMAURICE and LAIRD (1993) is unable to handle unequal cluster sizes adequately, as e.g. in family studies.

The authors recommend, based on the literature and their own experience, an application of the GEE only, if the number of clusters is at least 30 for a cluster size of about 4 for a low to moderate correlation. For high correlations between observations more independent clusters are necessary. Of course, the number of required clusters also depends on the number of explanatory variables. If the cluster size is large compared to the number of clusters, the GEE are probably not an appropriate analysing tool. In this situation random effect models or conditional models might be the better choice. When the number of clusters is small, careful modelling of the correlation needs to be done (PRENTICE, 1988). Also, one may want to use the bootstrap as discussed in MOULTON and ZEGER (1989) or SHERMAN and LE CESSIE (1996). We recommend to use IEE first and to model other association structures in a second step. To check for efficiency of the IEE, the findings of MANCL and LEROUX (1996) should be applied.

Acknowledgements

C. Kastner was supported by the Deutsche Forschungsgemeinschaft. The helpful comments of three anonymous reviewers are gratefully acknowledged.

References

AMOS, C. I., 1994: Robust variance-components approach for assessing genetic linkage in pedigrees. *American Journal of Human Genetics* **54**, 535–543.

AMOS, C. I., ZHU, D. K., and BOERWINKLE, E., 1996: Assessing genetic linkage and association with robust components of variance approaches. *Annals of Human Genetics* **60**, 143–160.

ANDOH, M. and UWOI, T., 1995: An interactive program of the GEE method for the analysis of longitudinal data. In: *SUGI 20 Proceedings*. SAS Institute, Inc., Cary, 1284–1289.

ARMINGER, G., 1992: Residuals and influential points in mean structures estimated with pseudo maximum likelihood methods. *Lecture Notes in Statistics* **78**, 20–26.

ARMINGER, G. and SOBEL, M. E., 1990: Pseudo-maximum likelihood estimation of mean and covariance structures with missing data. *Journal of the American Statistical Association* **85**, 195–203.

ASHBY, M., NEUHAUS, J. M., HAUCK, W. W., BACCHETTI, P., HEILBRON, D. C., JEWELL, N. P., SEGAL, M. R., and FUSARO, R. E., 1992: An annotated bibliography of methods for analysing correlated categorical data. *Statistics in Medicine* **11**, 67–99.

BIELER, G. S. and WILLIAMS, R. L., 1995: Cluster sampling techniques in quantal response teratology and developmental toxicity studies. *Biometrics* **51**, 764–776.

BINDER, D. A., 1983: On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* **51**, 279–292.

BISHOP, Y. M. M., FIENBERG, S. E., and HOLLAND, P. W., 1975: *Discrete multivariate analysis: Theory and practice*. MIT Press, Cambridge.

BOWMAN, D., CHEN, J. J., and GEORGE, E. O., 1995: Estimating variance functions in developmental toxicity studies. *Biometrics* **51**, 1523–1528.

BROWNE, M. W. and ARMINGER, G., 1995: Specification and estimation of mean- and covariance-structure models. In: G. Arminger, C. C. Clogg, and M. E. Sobel (Eds.): *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. Plenum, New York, 185–249.

CAREY, V., 1989: Data objects for matrix computations: An overview. *8th Proceedings of Computer Science and Statistics: 8th Annual Symposium on the Interface* **21**, 157–161.

CAREY, V., ZEGER, S. L., and DIGGLE, P., 1993: Modelling multivariate binary data with alternating logistic regression. *Biometrika* **80**, 517–526.

COCHRANE, D. and ORCUTT, G. H., 1949: Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association* **44**, 32–61.

CROWDER, M., 1985: Gaussian estimation for correlated binary data. *Journal of the Royal Statistical Society B* **47**, 229–237.

CROWDER, M., 1995: On the use of a working correlation matrix in using generalized linear models for repeated measurements. *Biometrika* **82**, 407–410.

DAVIS, C. S., 1991: Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials. *Statistics in Medicine* **10**, 1959–1980.

DAVIS, C. S., 1993: A computer program for regression analysis of repeated measures using generalized estimating equations. *Computer Methods and Programs in Biomedicine* **40**, 15–31.

DEMPSTER, A., LAIRD, N. M., and RUBIN, D. B., 1977: Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* **39**, 1–38.

DIEPGEN, T. L. and BLETTNER, M., 1996: Analysis of familial aggregation of atopic eczema and other atopic diseases by odds ratio regression models. *Journal of Investigative Dermatology* **106**, 977–981.

DIGGLE, P. J., 1992: Discussion of "Multivariate regression analysis for categorical data" by Liang, Zeger and Qaqish. *Journal of the Royal Statistical Society B* **54**, 28–29.

DIGGLE, P. J., LIANG, K. Y., and ZEGER, S. L., 1994: *Analysis of longitudinal data.* Oxford University Press, New York.

DUNCAN, T. E., DUNCAN, S. C., HOPS, H., and STOOLMILLER, M., 1995: An analysis of the relationship between parent and adolescent marijuana use via generalized estimating equation methodology. *Multivariate Behavioral Research* **30**, 317–339.

EMRICH, L. J. and PIEDMONTE, M. R., 1992: On some small sample properties of generalized estimating equations for multivariate dichotomous outcomes. *Journal of Statistical Computation and Simulation* **41**, 19–29.

FAHRMEIR, L. and TUTZ, G., 1994: *Multivariate statistical modelling based on generalized linear models.* Springer, New York.

FIRTH, D., 1992: Discussion of "Multivariate regression analysis for categorical data" by Liang, Zeger and Qaqish. *Journal of the Royal Statistical Society B* **54**, 24–26.

FIRTH, D., 1993: Recent developments in quasi-likelihood methods. *Proceedings of the ISI 49th Session*, Firenze, 341–358.

FITZMAURICE, G. M., 1995: A caveat concerning independence estimating equations with multivariate binary data. *Biometrics* **51**, 309–317.

FITZMAURICE, G. M. and LAIRD, N. M., 1993: A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**, 141–151.

FITZMAURICE, G. M. and LAIRD, N. M., 1995: Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association* **90**, 845–852.

FITZMAURICE, G. M., LAIRD, N. M., and ROTNITZKY, A., 1993: Regression models for discrete longitudinal responses. *Statistical Science* **8**, 284–309.

FITZMAURICE, G. M., LAIRD, N. M., and LIPSITZ, S. R., 1994: Analysing incomplete longitudinal binary responses: A likelihood-based approach. *Biometrics* **50**, 601–612.

FITZMAURICE, G. M., MOLENBERGHS, G., and LIPSITZ, S. R., 1995: Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society B* **57**, 691–704.

FRAMINGHAM, 1996: Familial aggregation and prevalence of myotopia in the Framingam Offspring Eye Study. The Framingham Offspring Eye Study group. *Archives of Ophthalmology* **114**, 326–332.

GANGE, S. J., LINTON, K. L. P., SCOTT, A. J., DEMETS, D. L., and KLEIN, R., 1994: A comparison of methods for correlated ordinal measures with ophthalmologic applications. *Statistics in Medicine* **14**, 1961–1974.

GLYNN, R. J. and ROSNER, B., 1994: Comparison of alternative regression models for paired binary data. *Statistics in Medicine* **13**, 1023–1036.

GOURIEROUX, C. and MONFORT, A., 1993: Pseudo-likelihood methods. In: *Handbook of Statistics*, Vol. 11, Eds. G. Maddala, C. R. Rao & H. Vinod, pp. 335–362. Amsterdam: Elsevier.

GOURIEROUX, C., MONFORT, A., and TROGNON, A., 1984: Pseudo maximum likelihood methods: Theory. *Econometrica* **52**, 682–700.

GREENE, W. H., 1993: *Econometric Analysis*, 2nd ed. MacMillan, New York.

GRÖMPING, U., 1993: *GEE: A SAS macro for longitudinal data analysis.* Technical Report, University of Dortmund, Department of Statistics.

GROVE, J. S., ZHAO, L. P., and QUIAOIT, F., 1993: Correlation analysis of twin data with repeated measures based on generalized estimating equations. *Genetic Epidemiology* **10**, 539–544.

GUNSOLLEY, J. C., GETCHELL, C., and CHINCHILLI, V. M., 1995: Small sample characteristics of generalized estimating equations. *Communications in Statistics – Computation and Simulation* **24**, 869–878.

HALL, C. B., ZEGER, S. L., and BANDEEN-ROCHE, K. J., 1994: *Added variable plots for regression with dependent data.* Technical Report, Department of Biostatistics, The Johns Hopkins University, Baltimore.

HALL, D. B. and SEVERINI, T. A., 1995: *Extended generalized estimating equations for clustered data.* Technical Report, University of Iowa.

HANSEN, L., 1982: Large sample properties of generalized methods of moments estimators. *Econometrica* **50**, 1029–1055.

HEAGERTY, P. J. and ZEGER, S. L., 1996: Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association* **91**, 1024–1036.

HUBER, P. J., 1967: The behaviour of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium*, 221–233.

JENNRICH, R. I. and SCHLUCHTER, M. D., 1986: Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42**, 805–829.

KARIM, M. and ZEGER, S. L., 1988: *GEE. A SAS macro for longitudinal data analysis.* Technical Report, Department of Biostatistics, The Johns Hopkins University, Baltimore, MD.

KASTNER, C., FIEGER, A., and HEUMANN, C., 1996: MAREG and WINMAREG – a tool for marginal regression models. *Statistical Software Newsletter in Computational Statistics and Data Analysis* **24**, 237–241.

KENWARD, M. G. and JONES, B., 1992: Alternative approaches to the analysis of binary and categorical repeated measurements. *Journal of Biopharmaceutical Statistics* **2**, 137–170.

KENWARD, M. G., LESAFFRE, E., and MOLENBERGHS, G., 1994a: Application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from likelihood in bivariate logistic regression. *Statistical Computation and Simulation* **44**, 133–148.

KENWARD, M. G., LESAFFRE, E., and MOLENBERGHS, G., 1994b: An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics* **50**, 945–953.

KENWARD, M. G. and SMITH, D. M., 1995a: Computing the generalized estimating equations with quadratic covariance estimation for repeated measurements. *Genstat Newsletter* **32**, 50–62.

KENWARD, M. G. and SMITH, D. M., 1995b: Computing the generalized estimating equations for repeated ordinal measurements. *Genstat Newsletter* **32**, 63–70.

KOCH, G. G., IMREY, P. B., and REINFURT, D. W., 1977: Linear model analysis of categorical data with incomplete response vectors. *Biometrics* **28**, 663–692.

LE CESSIE, S. and VAN HOUWELINGEN, J. C., 1994: Logistic regression for correlated binary data. *Applied Statistics* **43**, 95–108.

LEE, H. and STRAM, D. O., 1996: Segregation analysis of continuous phenotypes by using higher sample moments. *Genetic Epidemiology* **58**, 213–224.

LEE, A., SCOTT, A., and SOO, S., 1993a: Comparing Liang-Zeger estimates with maximum likelihood in bivariate logistic regression. *Communications in Statistics – Computation and Simulation* **44**, 133–148.

LEE, H., STRAM, D. O., and THOMAS, D. C., 1993b: A generalized estimating equations approach to fitting major gene models in segregation analysis of continuous phenotypes. *Genetic Epidemiology* **10**, 61–74.

LEISENRING, W., PEPE, M. S., and LONGTON, G., 1997: A marginal regression modelling framework for evaluating medical diagnostic tests. *Statistics in Medicine* **16**, 1263–1281.

LIANG, K. Y., 1992: Extensions of the generalized linear models in the past twenty years: Overview and some biomedical applications. *16th International Biometic Conference*, Hamilton, New Zealand, 27–38.

LIANG, K. Y. and BEATY, T. H., 1991: Measuring familial aggregation by using odds-ratio regression models. *Genetic Epidemiology* **8**, 361–370.

LIANG, K. Y. and PULVER, A. E., 1996: Analysis of case-control/family sampling design. *Genetic Epidemiology* **13**, 253–270.

LIANG, K. Y. and SELF, S. G., 1996: On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *Journal of the Royal Statistical Society* B **58**, 785–796.

LIANG, K. Y. and ZEGER, S. L., 1986: Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

LIANG, K. Y. and ZEGER, S. L., 1992: Regression analysis for correlated data. *Annual Review of Public Health* **14**, 43–68.

LIANG, K. Y., ZEGER, S. L., and QAQISH, B., 1992: Multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society B* **54**, 3–24.

LIPSITZ, S. R. and HARRINGTON, D. P., 1990: Analyzing correlated binary data using SAS. *Computers and Biomedical Research* **23**, 268–282.

LIPSITZ, S. R., LAIRD, N. M., and HARRINGTON, D. P., 1990: Using the jackknife to estimate the variance of regression estimators from repeated measures studies. *Communications in Statistics – Theory and Methods* **19**, 821–845.

LIPSITZ, S. R., LAIRD, N. M., and HARRINGTON, D. P., 1991: Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika* **78**, 153–160.

LIPSITZ, S. R., DEAR, K. B., and ZHAO, L. P., 1994a: Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics* **50**, 842–846.

LIPSITZ, S. M., FITZMAURICE, G., ORAV, E., and LAIRD, N. M., 1994b: Performance of generalized estimating equations in practical situations. *Biometrics* **50**, 270–278.

LIPSITZ, S. R., KIM, K., and ZHAO, L. P., 1994c: Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine* **14**, 1149–1163.

LIPSITZ, S. R., LAIRD, N. M., and HARRINGTON, D. P., 1994d: Weighted least squares analysis of repeated categorical measurements with outcomes subject to nonresponse. *Biometrics* **50**, 11–24.

LIPSITZ, S. R., FITZMAURICE, G. R., MOLENBERGHS, G., and Zhao, L. P., 1997: Quantile regression methods for longitudinal data with drop-outs: Application to CD4 cell counts of patients infected with the human immunodeficiency virus. *Applied Statistics* **46**, 463–476.

LIU, G. and LIANG, K. Y., 1997: Sample size calculations for studies with correlated observations. *Biometrics* **53**, 937–947.

LUENBERGER, D. G., 1984: *Linear and nonlinear programming*, 2nd ed. Addison-Wesley, Reading, Massachusetts.

LUMLEY, T., 1996a: Generalized estimating equations for ordinal data: A note on working correlation structures. *Biometrics* **52**, 354–361.

LUMLEY, T., 1996b: XLISP-Stat tools for building Generalised Estimating Equation models. *Journal of Statistical Software* **1**, 1–20.

MANCL, L. A. and LEROUX, B. G., 1996: Efficiency of regression estimates for clustered data. *Biometrics* **52**, 500–511.

MAY, W. L. and JOHNSON, W. D., 1995: Some applications of the analysis of multivariate normal data with missing observations. *Journal of Biopharmaceutical Statistics* **5**, 215–228.

MCCULLAGH, P. and NELDER, J., 1989: *Generalized linear models*, 2nd ed. London: Chapman & Hall.

MCDONALD, B. W., 1993: Estimating logistic regression parameters for bivariate binary data. *Journal of the Royal Statistical Society B* **55**, 391–397.

MILLER, M. E., 1995: Analysing categorical responses obtained from large clusters. *Applied Statistics* **44**, 173–186.

MILLER, M. E., DAVIS, C. S., and LANDIS, J. R., 1993: The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares. *Biometrics* **49**, 1033–1044.

MOULTON, L. H. and ZEGER, S. L., 1989: Analyzing repeated measures on generalized linear models via the bootstrap. *Biometrics* **45**, 381–394.

NEUHAUS, J. M., KALBFLEISCH, J. D., and HAUCK, W. W., 1991: A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* **59**, 25–35.

NEWEY, W. K., 1990: Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5**, 99–135.

NEWEY, W. K., 1993: Efficient estimation of models with conditional moment restrictions. In: G. Maddala, C. R. Rao, and H. Vinod (Eds.): *Handbook of Statistics*, Vol. 11. Elvsevier, Amsterdam, 419–454.

NORLEANS, M. X., 1995: *A generalized mixed linear model for the analysis of longitudinal data on an arbitrary scale.* Unpublished manuscript.

NUAMAH, I. F., QU, Y., and AMINI, S. B., 1996: A SAS macro for stepwise correlated binary regression. *Computer Methods and Programs in Biomedicine* **49**, 199–210.

O'HARA HINES, R. J., 1997a: Fitting generalized linear models to retrospectively sampled clusters with categorical responses. *Canadian Journal of Statistics* **25**, 159–174.

O'HARA HINES, R. J., 1997b: Analysis of clustered polytomous data using generalized estimating equations and working covariance structures. *Biometrics* **53**, 1552–1556.

O'HARA HINES, R. J., 1998: Comparison of two covariance structures in the analysis of clustered polytomous data using generalized estimating equations. *Biometrics*, in press.

OLSON, J. M., 1994a: Some empirical properties of an all-relative-pairs linkage test. *Genetic Epidemiology* **11**, 41–49.

OLSON, J. M., 1994b: Robust estimation of gene frequency and association parameters. *Biometrics* **50**, 665–674.

OLSON, J. M., 1995: Robust multipoint linkage analysis: An extension of the Haseman-Elston method. *Genetic Epidemiology* **12**, 177–193.

OLSON, J. M. and WIJSMAN, E. M., 1993: Linkage between quantitative trait and marker loci: Methods using all relative pairs. *Genetic Epidemiology* **10**, 87–102.

PAIK, M. C., 1988: Repeated measurement analysis for nonnormal data in small samples. *Communications in Statistics – Computation and Simulation* **17**, 1155–1171.

PAIK, M. C., 1992: Parametric variance function estimation for nonnormal repeated measurement data. *Biometrics* **48**, 19–30.

PAIK, M. C., 1997: The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association* **92**, 1320–1329.

PARK, T., 1993: A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. *Statistics in Medicine* **12**, 1723–1732.

PARK, T., 1994: Multivariate regression models for discrete and continuous repeated measurements. *Communications in Statistics – Theory and Methods* **23**, 1547–1564.

PARK, T. and SHIN, M. W., 1995: A practical extension of the generalized estimating equation approach for longitudinal data. *Communications in Statistics – Theory and Methods* **24**, 2561–2579.

PENDERGAST, J. F., GANGE, S. J., NEWTON, M. A., LINDSTROM, M. J., PALTA, M., and FISHER, M. R., 1996: A survey of methods for analyzing clustered binary response data. *International Statistical Review* **64**, 89–118.

PEPE, M. S. and ANDERSON, G. L., 1994: A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics – Simulation and Communication* **23**, 939–951.

PODGOR, M. J., HILLER, R., and THE FRAMINGHAM EYE STUDIES GROUP, 1996: Associations of types of lens opacities between and within eyes of individuals. *Statistics in Medicine* **15**, 145–156.

POLITO, J. M., REES, R. C., CHILDS, B., MENDELDOFF, A. I., HARRIS, M. L., and BAYLESS, T. M., 1996: Preliminary evidence for genetic anticipation in Crohn's disease. *Lancet* **23**, 798–800.

PREGIBON, D., 1983: An alternative covariance estimated for generalised linear models. *GLIM Newsletter* **13**, 51–55.

PREISSER, J. S. and QAQISH, B. F., 1996: Deletion diagnostics for generalised estimating equations. *Biometrika* **83**, 551–562.

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

PRENTICE, R. L. and ZHAO, L. P., 1991: Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47**, 825–839.

PRUSCHA, H., 1996: *Angewandte Methoden der Mathematischen Statistik*, 2nd ed. Stuttgart: Teubner.

QAQISH, B. F. and LIANG, K. Y., 1992: Marginal models for correlated binary responses with multiple classes and multiple levels of nesting. *Biometrics* **48**, 939–950.

Q.G.E., 1994: *EE: Estimating Equations.* Technical Report, Fred Hutchinson Cancer Research Center, Quantitative Genetic Epidemiology.

QU, Y., WILLIAMS, G. W., BECK, G. J., and MEDENDORP, S. V., 1992: Latent variable models for clustered dichotomous data with multiple subclusters. *Biometrics* **48**, 1095–1102.

ROBINS, J. M. and ROTNITZKY, A., 1995: Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* **90**, 122–129.

ROBINS, J. M., ROTNITZKY, A., and ZHAO, L. P., 1994: Estimation of regression coefficients when a regressor is not always observed. *Journal of the American Statistical Association* **89**, 846–866.

ROBINS, J. M., ROTNITZKY, A., and ZHAO, L. P., 1995: Analysis of semiparametric regression models for repeated outcomes under the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.

ROTNITZKY, A., 1988: *Analysis of generalized linear models for cluster correlated data*. PhD thesis, University of California, Berkeley.

ROTNITZKY, A. and JEWELL, N., 1990: Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77**, 485–497.

ROTNITZKY, A. and ROBINS, J. M., 1995a: Semi-parametric estimation of models for means and covariances in the presence of missing data. *Scandinavian Journal of Statistics* **22**, 323–333.

ROTNITZKY, A. and ROBINS, J. M., 1995b: Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* **82**, 805–820.

ROTNITZKY, A. and ROBINS, J. M., 1997: Analysis of semiparametric regression models with non-ignorable non-response. *Statistics in Medicine* **16**, 81–102.

ROTNITZKY, A. and WYPIJ, D., 1994: A note on the bias of estimators with missing data. *Biometrics* **50**, 1163–1170.

ROYALL, R. M., 1986: Model robust confidence intervals using maximum likelihood estimation. *International Statistical Review* **54**, 221–226.

RUBIN, D. B., 1976: Inference and missing data. *Biometrika* **63**, 581–592.

RYAN, L., 1992: The use of generalized estimating equations for risk assessment in developmental toxicity. *Risk Analysis* **12**, 439–447.

SCHNEIDER, C., KOCH, M. C., REINERS, K., ZIEGLER, A., REIMERS, C. D., MEINCK, H.-M., BROICH, P., GONSCHOREK, A. S., TOYKA, K. V., and RICKER, K., 1998: Anticipation in Proximal Myotonic Myopathy (PROMM): A Study in 80 Families. Submitted to *Brain*.

SHARPLES, K. and BRESLOW, N., 1992: Regression analysis of correlated binary data: Some small sample results for estimating equations. *Statistical Computation and Simulation* **42**, 1–20.

SHERMAN, M. and LE CESSIE, S., 1997: A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics – Simulation and Communication* **26**, 901–925.

SHIH, M., 1997: Sample size and power calculations for periodontal and other studies with clustered samples using the method of generalized estimating equations. *Biometrical Journal* **39**, 899 to 908.

SPIESS, M. and HAMERLE, A., 1996: On the properties of GEE estimators in the presence of invariant covariates. *Biometrical Journal* **38**, 931–940.

STRAM, D. O., WEI, L., and WARE, J., 1988: Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariables. *Journal of the American Statistical Association* **83**, 631–637.

STRAM, D. O., LEE, H., and THOMAS, D. C., 1993: Use of generalized estimating equations in segregation analysis of continuous outcomes. *Genetic Epidemiology* **10**, 575–579.

TAN, M., QU, Y., and KUTNER, M. H., 1997: Model diagnostics for marginal regression analysis of correlated binary data. *Communications in Statistics – Simulation and Communication* **26**, 539–558.

TEN HAVE, T. R. and CHINCHILLI, V. M., 1995: The analysis of bioequivalence with respect to TMAX under a 2×2 crossover design. *Journal of Biopharmaceutical Statistics* **5**, 185–199.

TOSTESON, T. D., ROSNER, B., and REDLINE, S., 1991: Logistic regression for clustered binary data in proband studies with application to familial aggregation of sleep disorders. *Biometrics* **47**, 1257–1265.

TROXEL, A. B., LIPSITZ, S. R., and BRENNAN, T. A., 1997: Weighted estimating equations with nonignorably missing response data. *Biometrics* **53**, 857–869.

TRÉGOUËT, D. A., DUCIMETIÈRE, P., and TIRET, L., 1997: Testing association between candidate-gene markers and phenotype in related individuals, by use of estimating equations. *American Journal of Human Genetics* **61**, 189–199.

WEI, L. and STRAM, D., 1988: Analysing repeated measurements with possibly missing observations by modelling marginal distributions. *Statistics in Medicine* **7**, 139–148.

WHITE, H., 1982: Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

WHITTEMORE, A. S., 1995: Logistic regression of family data from case-control studies. *Biometrika* **82**, 57–67.

WHITTEMORE, A. S. and GONG, G., 1994: Segregation analysis of case-control data using generalized estimating equations. *Biometrics* **50**, 1073–1087.

WILLIAMSON, J. M., KIM, K., and LIPSITZ, S. R., 1995: Analyzing bivariate ordinal data using a global odds ratio. *Journal of the American Statistical Association* **90**, 1432–1437.

XIE, F. and PAIK, M. C., 1997: Generalized estimating equation model for binary outcomes with missing covariates. *Biometrics* **53**, 1458–1466.

XIE, F. and PAIK, M. C., 1997: Multiple imputation methods for the missing covariates in generalized estimating equation. *Biometrics* **53**, 1538–1546.

ZEGER, S. L., 1988: Commentary. *Statistics in Medicine* **7**, 95–107.

ZEGER, S. L. and LIANG, K. Y., 1986: Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–130.

ZEGER, S. L. and LIANG, K. Y., 1992: An overview of methods for the analysis of longitudinal data. *Statistics in Medicine* **11**, 1825–1839.

ZEGER, S. L., LIANG, K. Y., and SELF, S. G., 1985: The analysis of binary longitudinal data with time-independent covariates. *Biometrika* **72**, 31–38.

ZHAO, L. P., 1994: Segregation analysis of human pedigrees using estimating equations. *Biometrika* **81**, 197–209.

ZHAO, L. P. and LE MARCHAND, L., 1992: An analytical method for assessing patterns of familial aggregation in case-control studies. *Genetic Epidemiology* **9**, 141–154.

ZHAO, L. P. and PRENTICE, R. L., 1990: Correlated binary regression using a generalized quadratic model. *Biometrika* **77**, 642–648.

ZHAO, L. P. and PRENTICE, R. L., 1991: Use of a quadratic exponential model to generate estimating equations for means, variances, and covariances. In: V. P. Godambe (Ed.): *Estimating Functions*. Oxford University Press, Oxford, 103–117.

ZHAO, L. P., PRENTICE, R. L., and SELF, S. G., 1992: Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society B* **54**, 805–811.

ZHAO, L. P., HOLTE, S., CHEN, Y., QUIAOIT, F., and PRENTICE, R. L., 1996a: *Aggregation analysis of family data from case-control studies*. Technical Report, Fred Hutchinson Cancer Research Center, Seattle.

ZHAO, L. P., LIPSITZ, S. R., and LEW, D., 1996b: Regression analysis with missing covariate data using estimating equations. *Biometrics* **52**, 1165–1182.

ZHU, Y., KREWSKI, D.; and ROSS, W. H., 1994: Dose-response models for correlated multinomial data from developmental toxicity studies. *Applied Statistics* **43**, 583–598.

ZIEGLER, A., 1994: *Verallgemeinerte Schätzgleichungen zur Analyse korrelierter Daten*. PhD thesis, University of Dortmund, Department of Statistics.

ZIEGLER, A., 1995: The different parameterizations of the GEE1 and the GEE2. *Lecture Notes in Statistics* **104**, 315–324.

ZIEGLER, A., 1997: Practical considerations of the jackknife estimator of variance for generalized estimating equations. *Statistical Papers* **38**, 363–369.

ZIEGLER, A. and ARMINGER, G., 1995: Analyzing the employment status with panel data from the GSOEP – a comparison of the MECOSA and the GEE1 approach for marginal models. *Vierteljahreshefte zur Wirtschaftsforschung* **64**, 72–80.

ZIEGLER, A. and ARMINGER, G., 1996: Parameter estimation and regression diagnostics using generalized estimating equations. In: F. Faulbaum and W. Bandilla (Eds.): *SoftStat '95, Advances in Statistical Software 5*. Lucius & Lucius, Stuttgart, 229–237.

ZIEGLER, A. and GRÖMPING, U., 1998: Generalized estimating equations in commercial statistical software packages. *Biometrical Journal* **40**, 247–262.

ZIEGLER, A. and KASTNER, C., 1997: A minimum distance estimation approach to estimate the recombination fraction from a marker locus in robust linkage analysis for quantitative traits. *Biometrical Journal* **39**, 765–775.

ZIEGLER, A., BLETTNER, M., KASTNER, C., and CHANG-CLAUDE, J., 1998: Identifying influential families using regression diagnostics for Generalized Estimating Equations. *Genetic Epidemiology*, in press.

ANDREAS ZIEGLER
Medical Centre for Methodology and Health Research
Institute of Medical Biometry and Epidemiology
Philipps-University of Marburg
Bunsenstr. 3
35033 Marburg
Germany
phone no.: ++49/64 21/28-57 87
fax: ++49/64 21/28-89 21
e-mail: ziegler@mailer.uni-marburg.de

CHRISTIAN KASTNER
Institute of Statistics
LMU München
Ludwigstr. 33
80539 München
Germany
e-mail: kchris@stat.uni-muenchen.de

MARIA BLETTNER
International Agency for Research on Cancer
150, cours Albert-Thomas
69372 Lyon Cedex 08
France
e-mail: blettner@iarc.fr

Received, October 1997
Revised, December 1997
Accepted, March 1998