2011

Generalized Estimating Equations and Gaussian Estimation in Longitudinal Data Analysis

Xuemao Zhang
University of Windsor

Generalized Estimating Equations and Gaussian Estimation in Longitudinal Data Analysis

by

Xuemao Zhang

A Dissertation

Submitted to the Faculty of Graduate Studies

through the Department of Mathematics and Statistics

in Partial Fulfillment of the Requirements for

the Degree of Doctor of Philosophy at the

University of Windsor

Windsor, Ontario, Canada

2011

GEE and Gaussian Estimation in Longitudinal Data Analysis

by
Xuemao Zhang

APPROVED BY:

Dr. P. Song, External Examiner
University of Michigan

Dr. Y. Aneja
Odette School of Business

Dr. A. A. Hussein
Department of Mathematics and Statistics

Dr. M. Hlynka
Department of Mathematics and Statistics

Dr. S. R. Paul, Advisor
Department of Mathematics and Statistics

Dr. K. Taylor, Chair of Defense
Department of Chemistry and Biochemistry

June 22, 2011

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication. I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix. I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

In this dissertation, we first develop a Gaussian estimation procedure for the estimation of regression parameters in correlated (longitudinal) binary response data using working correlation matrix and compare this method with the GEE (generalized estimating equations) method and the weighted GEE method. A Newton-Raphson algorithm is derived for estimating the regression parameters from the Gaussian likelihood estimating equations for known correlation parameters. The correlation parameters of the working correlation matrix are estimated by the method of moments. Consistency properties of the estimators are discussed. A simulation comparison of efficiency of the Gaussian estimates and the GEE estimates of the regression parameters shows that the Gaussian estimates using the unstructured correlation matrix of the responses for a subject are, in general, more efficient than those by the other methods compared. The next best are the Gaussian estimates using the general autocorrelation structure. Two data sets are analyzed and a discussion is given.

The main advantage of GEE is its asymptotic unbiased estimation of the marginal regression coefficients even if the correlation structure is misspecified. However, the technique requires that the sample size should be large. In this dissertation, two bias corrected GEE estimators of the regression parameters in longitudinal data are proposed when the sample size is small. Simulations show that the proposed methods do well in reducing bias and have, in general, higher efficiency than the GEE estimates. Two examples are analyzed and a discussion is given.

The current GEE method focuses on the modeling of the working correlation matrix assuming a known variance function. However, Wang and Lin (2005) showed that

if the variance function is misspecified, the correct choice of the correlation structure may not necessarily improve estimation efficiency for the regression parameters. In this dissertation, we propose a GEE approach to estimate the variance parameters when the form of the variance function is known. This estimation approach borrows the idea of Davidian and Carroll (1987) by solving a non-linear regression problem where residuals are regarded as the responses and the variance function is regarded as the regression function. Simulations show that the proposed method performs as well as the modified pseudolikelihood approach developed by Wang and Zhao (2007).

Dedication

This thesis is dedicated to my wife, Yuxia Niu. I thank her for her love and support throughout the years. It is also dedicated to my parents who have been a constant source of encouragement.

Acknowledgements

I would like to express my profound gratitude to my supervisor Dr. Paul. He never hesitated to provide me assistance when I need help throughout my study. The doctoral program under his supervision has prepared me well for my future professional career. This dissertation could not have been accomplished without his insights into all the statistical subjects. He also has made numerous very useful suggestions to the thesis composition including wording and grammar. Moreover, I am very grateful to Dr. Paul for the Research Assistantship he has provided to me.

I would like to thank Dr. Hlynka and Dr. Hussein as the department readers. Their remarks have made the thesis more rigorous and readable. I am also grateful to Dr. Aneja of the Department of Management Science, Odette School of Business, University of Windsor and Dr. Song of University of Michigan for their valuable comments and suggestions.

I would like to thank the University of Windsor for providing me a Graduate Assistantship, and the Ontario Ministry of Training, Colleges and Universities for providing me an Ontario Graduate Scholarship during my graduate study. These financial supports have enabled me to finish the doctoral program more easily.

Last but not least, I wish to thank my wife and my parents for their constant love, encouragement and support. They have always been eager to help me in any way they could.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Longitudinal data arise in many fields such as biomedical and social sciences. Longitudinal data are characterized by repeated measurements taken on each of a number of subjects over time. In these studies it is reasonable to assume that the subjects are independent, but the repeated measurements taken on each subject may not be uncorrelated. The purpose of longitudinal data analysis is to model the relationship of the repeated measurements of each subject to the associated covariates. As an example consider the data from the Six Cities study, a longitudinal study of the health effects of air pollution that was analyzed by Fitzmaurice and Laird (1993). The data set contains complete records on 537 children from Steubenville, Ohio, each of whom was examined annually at ages 7 through 10. The repeated binary response is the wheezing status (1=yes, 0=no) of a child at each occasion. The purpose of the study is to model the probability of the wheezing status as a function of the child's age, his/her mother's maternal smoking habit (a binary variable MS with 1 if the mother smoked regularly and 0 otherwise) and their interactions.

There are three types of models for longitudinal data analysis: transition or fully-conditional models (Korn and Whittemore, 1979, Rosner, 1984 and Zeger and Qaqish, 1988 etc.), random-effects models (Rao, 1965, Laird and Ware, 1982 and Stiratelli, Laird, and Ware, 1984 etc.) and marginal models (Liang and Zeger, 1986, Zeger and Liang, 1986 and Prentice and Zhao, 1991 etc.). Transition models are used to specify the conditional distribution of each response given the past responses. Random-effects models describe the natural heterogeneity among subjects. Marginal models are used to characterize the marginal expected value of a subject's response as a function of the subject's covariates. Diggle, Liang and Zeger (1994) discussed these models in detail. The study of the marginal model of longitudinal data analysis is our focus in this thesis.

The complication of longitudinal data analysis is partly due to the lack of a rich class of models such as the multivariate normal for the joint distribution of the responses of a subject. Therefore, a robust method that avoids full distributional assumptions of the likelihood approach is required. One such method is the generalized estimating equations (GEE) approach proposed by Liang and Zeger (1986) and Zeger and Liang (1986). The GEE method is developed from the theory of generalized linear models (GLM) by Nelder and Wedderburn (1972) and optimal inference functions established by Godambe (1960).

GEE is used to estimate the regression parameters in marginal models of longitudinal data in which the link function and variance function take the forms of those in GLM. GLM extends the classical linear models in two aspects. First, the response variables are from an exponential family which includes the normal distribution as a special case. Second, the monotone link function which relates the expected responses and the linear predictor may not be linear. Wedderburn (1974) developed the quasi-likelihood method in which only the first two moments, mean and variance, are specified for estimating the regression parameters when the distribution may not be from an exponential family. That is, the quasi-likelihood method does not assume a full distributional specification. McCullagh (1983) extended the quasi-likelihood method to multivariate cases where the components of a response vector are independent. GEE is a further extension of the quasi-likelihood method for the analysis of longitudinal data. It uses the working correlation matrix to take into account the correlation between the repeated measurements of a subject. The GEE estimator for the regression parameter β is consistent even if the working correlation structure is misspecified. However, correct specification of the correlation structure can improve the estimation efficiency of the regression parameter (Wang and Carey, 2003).

The working correlation structure in GEE is not fully understood though there is a lot of literature on this subject. There is a pitfall in estimating the correlation parameters. Crowder (1995) found that in some cases the parameters involved in the working correlation matrix are subject to an uncertainty of definition which can lead to a breakdown of asymptotic properties of the estimators (see also Crowder, 2001). Further, the misspecification of the correlation structure can result in loss of efficiency of the regression parameters (Wang and Carey, 2003). In fact, there is a history of controversy over choosing the working correlation structure $R(\rho)$ in GEE to obtain high efficiency of the estimators of the regression parameters. Sutradhar and Das (1999) considered a binary logistic regression model with cluster level covariates and showed by simulations that in many cases of misspecification of working correlation structures, the independence GEE approach yields more efficient estimators. In a subsequent paper Sutradhar and Das (2000) found, again by simulations, that if the model includes within-cluster covariates then the independence GEE approach yields less efficient estimators. Wang and Carey (2003) showed that the choice of working correlation structure $R(\rho)$ has a substantial impact on the efficiency of regression parameter estimators. The reason that the independence GEE performs well is that the design of Sutradhar and Das (1999) is balanced in the sense that the covariate pattern is the same for all individuals.

Gaussian estimation introduced by Whittle (1961) is another estimation technique with no distributional assumptions. It uses the normal log-likelihood as the estimation function without assuming that the data are normally distributed. The Gaussian estimation procedure has been shown to have good properties in a number of applications. For example, Crowder (1985) showed by simulation that a Gaussian estimate of the correlation parameter of equi-correlated clustered binary data has high efficiency. Paul and Islam (1998), again, by simulation, showed that a Gaussian estimator of the overdispersion parameter in clustered binomial data has best efficiency in comparison to likelihood, quasi-likelihood and extended quasi-likelihood estimates. Wang and Zhao (2007) used Gaussian estimation for the analysis of longitudinal data when the covariance function is modelled by additional variance parameters to the mean parameters.

The GEE technique is asymptotic. Thus, in the case of small sample sizes, GEE may result in biased estimates. Notice that the GEE function is an extension of the quasi-likelihood which is the true likelihood when the distribution is from an exponential family. This motivates us to use the bias-correction technique in maximum likelihood estimation to reduce the bias. Under general conditions, maximum likelihood (ML) estimators are consistent. But they are not unbiased generally. Cox and Snell (1968) provided general results for the first-order correction of bias of ML estimators for any distribution. Cordeiro and Klein (1994) gave a general matrix formula for computing the bias of the ML estimates. Firth (1993) showed that the order $1/n$ bias of the ML estimator can be removed by introducing an appropriate bias term into the score function. Now, if the score function in GEE is regarded as a true likelihood, then the bias reduction for the maximum likelihood method can be applied.

For the working covariance matrix in GEE, the current method focuses only on the modeling of the working correlation $R(\rho)$ and the variance function v is treated as known which is of the form in GLM, a function of the marginal mean μ. However, in practice the distribution of the data may not be from a GLM and we tend to choose a wrong variance function. Wang and Lin (2005) investigated the impacts of misspecifying the variance function on estimators of the regression parameters. They show that if the variance function is misspecified, the scorrect choice of the correlation structure may not necessarily improve estimation efficiency. This can be understood from the logic that modeling of the correlation structure is based on the correct modeling of the variance function. The best choice of the working correlation may no longer be the true one for estimating β if the specified variance function is far from the true one (Wang and Zhao, 2007). Therefore, the variance function plays a more important role than the correlation structure.

In this dissertation, we deal with three problems. We first explore Gaussian estimation for longitudinal data to improve estimation efficiency. Second, we propose two bias correction procedures to reduce the biases of GEE estimates of regression parameters when the sample size is small. Last, we investigate how the variance functions affect the estimation efficiency and propose a GEE approach to estimate the additional variance parameters in the variance function to improve estimation efficiency. This estimation approach borrows an idea of Davidian and Carroll (1987) by solving a non-linear regression problem where residuals are regarded as the responses and the variance function is regarded as the regression function.

In Chapter 2, we do a literature review which covers some definitions and rules in matrix calculus, generalized linear models (GLM), quasi-likelihood method, generalized estimating equations (GEE) and Gaussian copula regression models.

In Chapter 3, we study Gaussian estimation for longitudinal binary data. The purpose of this chapter is to develop and investigate the Gaussian estimation procedure for the estimation of regression parameters in longitudinal binary response data and compare this method with the GEE and related methods. As in the GEE we use a working correlation matrix for the responses of each individual. Consistency of the estimates of the regression parameters is ensured by carefully choosing a robust working correlation structure: general autocorrelation or unstructured. Efficiencies of the estimates are then compared with the GEE method and the weighted GEE approach by Chaganty and Joe (2004).

In Chapter 4, we study bias-correction in GEE estimation. By treating the GEE function as a likelihood score function, we apply the bias correction technique of Cordeiro and Klein (1994) and Firth (1993). The former method is corrective. That is, the GEE estimator is first calculated then corrected. The latter method is preventive in which a bias term is introduced into the GEE function.

In Chapter 5, we focus on the study of effects of variance function on the estimation efficiency of the regression parameters. The variance parameters in the variance function of known form are estimated using a GEE approach by solving a non-linear regression problem by regarding the residuals as responses and the variance function as regression function. This idea of the estimation method is borrowed from Davidian and Carroll (1987) in the setting of heteroscedastic regression models. Our proposed method is then compared with the pseudolikelihood approach by Wang and Zhao (2007).

In the last chapter, we summarize the findings of the dissertation and conclude with a related future research subject. When the covariates are time-dependent, the marginal regression analysis using GEE methods usually results in biased estimates of regression parameters. Future research will introduce a proper bias term into the GEE function or choose an appropriate weight in GEE to reduce the bias resulting from the time-dependent covariates.

CHAPTER 2

Literature Review

2.1. Definitions and rules in matrix calculus

In this section, we review the definition of the derivative of a function and the chain rule, product rule and the Kronecker product rule in matrix calculus (Magnus and Neudecker, 1988).

The vec operator and Kronecker products are used frequently in matrix calculus. The vec operator vectorizes a matrix by stacking the columns of the matrix one under the other. Let A be an $m \times n$ matrix and B a $p \times q$ matrix. The Kronecker product $A \otimes B$ of A and B which is $mp \times nq$ dimensional is defined by

$$\begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}.$$

Let f be a scalar function of an $n \times 1$ vector x. The derivative of f is defined as

$$Df(x) = (D_1 f(x), \ldots, D_n f(x)) = \frac{\partial f(x)}{\partial x},$$

where $D_j f$ is the derivative of f with respect to the jth variable, holding the other variables fixed. If f is an $m \times 1$ vector function of x, then the derivative (or *Jacobian matrix*) of f is the $m \times n$ matrix

$$Df(x) = \frac{\partial f(x)}{\partial x} \text{ with } [Df]_{ij} = \frac{\partial f_i(x)}{\partial x_j}.$$

Defining derivatives of matrices with respect to matrices is accomplished by vectorizing the matrices.

Definition 2.1.1. *Let F be a differentiable $m \times p$ real matrix function of an $n \times q$ matrix of real variables X. The* Jacobian matrix *of F at X is the $mp \times nq$ matrix*

$$DF(X) = \frac{\partial vecF(X)}{\partial(vecX)}.$$

A general *product rule* is defined as follows. Suppose that $f : \mathbb{R}^n \to \mathbb{R}^{m\times p}$ and $g : \mathbb{R}^n \to \mathbb{R}^{p\times q}$, so $f(x)g(x) : \mathbb{R}^n \to \mathbb{R}^{m\times q}$. Then

$$Df(x)g(x) = (g(x)^T \otimes I_m)Df(x) + (I_q \otimes f(x))Dg(x). \tag{2.1.1}$$

Furthermore, the *chain rule* involves matrix multiplication, which requires conformability. Given two functions $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^p \to \mathbb{R}^m$, the derivative of the composite function is

$$D[f(g(x))] = f'(g(x))g'(x).$$

Definition 2.1.2. *Let A be an $m \times n$ matrix. The vectors vec A and vec A^T clearly contain the same mn components, but in a different order. Hence there exists a unique $mn \times mn$ permutation matrix which transforms vec A into vec A^T. This matrix is called the* commutation matrix *and is denoted by K_{mn}. If $m = n$, K_{nn} is often written as K_n.*

Theorem 2.1.1. *Let $U : S \to \mathbb{R}^{n\times q}$ and $V : S \to \mathbb{R}^{p\times r}$ be two matrix functions defined and differentiable on an open set S in $\mathbb{R}^{m\times s}$. Then the Kronecker product $U \otimes V$ is differentiable on S and*

$$D(U \otimes V) = (I_q \otimes K_{r,n} \otimes I_p)\big[(I_{nq} \otimes vecV)DU + (vecU \otimes I_{pr})DV\big]. \tag{2.1.2}$$

2.2. Generalized linear models

The generalized linear model (GLM) developed by Nelder and Wedderburn (1972) is a generalization of normal linear models. It requires that the response variables be from an exponential family and the expected responses be a function of the linear predictors.

For the scalar observation z, suppose the probability density function is given by

$$f_Z(z;\theta,\phi) = \exp\{[z\theta - b(\theta)]/a(\phi) + c(z,\phi)\} \tag{2.2.1}$$

for some functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. This is called an exponential family with canonical parameter θ if ϕ is known. It can be seen that $\mathrm{E}(Z) = b^{'}(\theta)$. Moreover, the variance of Z is related to its expected value by $\mathrm{Var}(Z) = b^{''}(\theta)a(\phi)$, where $b^{''}(\theta)$ is called the variance function and ϕ is called the dispersion parameter.

Let Y and μ be $n \times 1$ dimensional vectors. The classical linear model can be rearranged to the following tripartite form (see McCullagh and Nelder, 1983):

1. The random component: Y has independent Normal distribution with constant variance σ^2 and $\mathrm{E}(Y) = \mu$.
2. The systematic component: covariates in the form of an $n \times p$ design matrix $X = (x_1^T, x_2^T, \ldots, x_n^T)^T$ produce a linear predictor η given by

   $$\eta = X\beta,$$

   where β is a $p \times 1$ regression parameter vector.
3. The link between the random and systematic components is given by

   $$\mu = \eta.$$

Generalized linear models generalize the classical linear models by allowing two extensions. First, the distribution in part 1 comes from an exponential family which includes the normal distribution as a special case. Secondly, the link between the random and systematic components is given by $\eta = g(\mu)$, where g is called the link function which is monotone and differentiable.

2.3. Quasi-likelihood

The quasi-likelihood method proposed by Wedderburn (1974) does not depend on the specification of a full distribution, such as a density function from an exponential family. Instead it just requires the structure of the mean and variance, that is, the first two moments. Moreover, the variance generally is a function of the expected value.

Let $y_1, \ldots, y_n$ be independent responses with means $\mathrm{E}(y_i) = \mu_i$ and variance $\mathrm{Var}(y_i) = \phi v(\mu_i)$, where μ_i is a function of unknown regression parameters $\beta_1, \beta_2, \ldots, \beta_p$, $v(\cdot)$ is a known variance function and ϕ is a scalar or dispersion parameter. Then the quasi-likelihood of a single observation y_i is given by

$$Q(\mu_i; y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi v(t)}\, dt. \tag{2.3.1}$$

And the quasi-likelihood for the complete data is given by the sum of the individual contributions

$$Q(\mu; y) = \sum_{i=1}^{n} Q(\mu_i; y_i).$$

The estimates of the regression parameters are obtained by solving a set of score-like equations

$$S_k(\beta) = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_k} (v(\mu_i))^{-1} (y_i - \mu_i) = 0, \quad k = 1, \ldots, p.$$

Let $y = (y_1, \ldots, y_n)^T$ and $\mu = (\mu_1, \ldots, \mu_n)^T$. Then, in matrix notation, these equations can be written as

$$S(\beta) = (\partial \mu / \partial \beta)^T V^{-1} (y - \mu) = 0, \tag{2.3.2}$$

where $V = \mathrm{diag}\{v(\mu_i)\}$.

For simple models, such as Normal distribution and Poisson distribution, the log likelihood score and the quasi-likelihood function are identical.

2.4. Generalized estimating equations

Generalized estimating equations (GEE) (Liang and Zeger 1986, Zeger and Liang 1986) generalize the quasi-likelihood method to analyze longitudinal/clustered data. In a marginal model, the analyst is interested in modeling the marginal expectation as a function of explanatory variables. The GEEs are used to characterize the marginal expectation of a set of outcomes as a function of a set of covariates.

We illustrate the longitudinal data framework as follows. The clustered data can be described in a similar way. Let $y_i = (y_{i1}, \dots, y_{in_i})'$ be the response vector for the ith subject, $i = 1, \dots, N$. Assume the N subjects are independent while the repeated measurements taken on each subject are correlated. Associated with each measurement y_{ij} is a vector of covariates $x_{ij} = (x_{ij1}, \dots, x_{ijp})'$, $j = 1, \dots, n_i, i = 1, \dots, N$. Let $X_i = (x_{i1}, \dots, x_{in_i})'$ be the $n_i \times p$ design matrix for the ith subject. Define μ_i be the expectation of y_i and suppose that

$$\mu_i = h(X_i\beta),$$

where β is a $p \times 1$ vector of regression parameters of interest and the inverse of h is referred as the link function. Also assume that the variance of y_{ij} is expressed as a known monotone function, v, of μ_{ij},

$$\mathrm{Var}(y_{ij}) = \phi v(\mu_{ij}),\ j = 1, \dots, n_i,$$

where ϕ is a dispersion parameter.

GEE method uses a common working correlation matrix for the longitudinal responses for each subject. The word “working” means that the correlation structure may not be correctly specified. Let $R(\rho)$ be a working correlation matrix completely specified by the parameter vector ρ of length q. Then $\phi W_i = \phi A_i^{1/2} R(\rho) A_i^{1/2}$ is the corresponding working covariance matrix, where $A_i = \text{diag}\{v(\mu_{ij})\}$ is a diagonal matrix, $i = 1, \dots, N$. For given consistent estimates of ϕ and ρ, the estimate $\hat{\beta}$ is the solution of the GEE equations

$$\sum_{i=1}^{N} D_i^T W_i^{-1} (y_i - \mu_i) = 0, \tag{2.4.1}$$

where $D_i = \frac{\partial \mu_i}{\partial \beta}$. It can be seen that if $R(\rho)$ is an identity matrix, then the GEE equations are quasi-likelihood estimating equations.

The estimator $\hat{\beta}$ of β obtained by solving GEE equation (2.4.1) is consistent even if the correlation structure is misspecified. However, the misspecification of the correlation structure may result in inefficient estimates of the regression parameters (for more details, see Wang and Carey, 2003).

Given consistent estimates $\hat{\rho}$ and $\hat{\phi}$ of the correlation and dispersion parameters, under mild regularity conditions (the parameter space is an open set; the GEE function $\sum_{i=1}^{N} D_i^T W_i^{-1}(y_i - \mu_i)$ is continuously differentiable; $|\partial \hat{\rho}(\beta, \phi)/\partial \phi| \le O_p(1)$), $N^{1/2}(\hat{\beta} - \beta)$ is asymptotically multivariate normal with mean zero and sandwich covariance matrix

$$\lim_{N \to \infty} N \left(\sum_{i=1}^{N} D_i^T W_i^{-1} D_i \right)^{-1} \left[\sum_{i=1}^{N} D_i^T W_i^{-1} \operatorname{Cov}(Y_i) W_i^{-1} D_i \right] \left(\sum_{i=1}^{N} D_i^T W_i^{-1} D_i \right)^{-1}.$$

This covariance matrix is estimated by replacing $\hat{\beta}$ by β and $\operatorname{Cov}(Y_i)$ by its estimated covariance matrix $(y_i - \hat{\mu}_i)^T (y_i - \hat{\mu}_i)$. The resulting estimated asymptotic covariance matrix is called the robust covariance matrix estimator of $\hat{\beta}$.

Instead of using the moment estimates of ϕ and ρ in generalized estimating equations, when analyzing correlated binary reponses, Prentice (1988) simultaneously modeled the mean and correlation profiles. In the estimation, a second set of GEEs to estimate the correlation parameters is added. The moment estimating equations for ρ is given by

$$u(\rho) = \sum_{i=1}^{N} \mathbf{A}_i{}^T \mathbf{H}_i{}^{-1}[\pi_i - \nu_i(\beta, \rho)] = 0, \tag{2.4.2}$$

where $\pi_i = \{\pi_{i12}, \pi_{i13}, \cdots, \pi_{i23}, \cdots\}, \nu_i = \{\nu_{i12}, \nu_{i13}, \cdots, \nu_{i23}, \cdots\}, \pi_{ist} = y_{is}y_{it}/(p_{is}(1-p_{is})p_{it}(1-p_{it}))^{1/2}, \nu_{ist} = \mathrm{E}(\pi_{ist}|x_i)$ for $s < t$, $\mathbf{A}_i = \partial\nu_i/\partial\rho$ and $\mathbf{H}_i = \mathrm{diag}(\mathrm{Var}(\pi_i))$. This simple estimating equation approach for means and covariances applies similarly to other types of response variables than binary. Prentice (1988) also established asymptotic normality for the joint distribution of his estimates of β and ρ.

Lipsitz, Laird and Harrington (1991) modified the estimating equations of Prentice and modeled the association between binary responses based on odds ratios. This approach is useful if the odds ratio are of interest themselves, and not confined between $(-1, 1)$.

Prentice and Zhao (1991), extending the idea of Prentice (1988), introduced estimating equations (GEE2) in an ad hoc fashion for means and covariances. The GEE2 can be written as

$$N^{-1/2} \sum_{i=1}^{N} D_i^T V_i^{-1} f_i = 0, \tag{2.4.3}$$

where $D_i = \begin{pmatrix} \partial^T\mu_i/\partial\beta & 0 \\ \partial^T\sigma_i/\partial\beta & \partial^T\sigma_i/\partial\rho \end{pmatrix}$, $V_i = \begin{pmatrix} \mathrm{Var}(y_i) & \mathrm{Cov}(y_i, s_i) \\ \mathrm{Cov}(s_i, y_i) & \mathrm{Var}(s_i) \end{pmatrix}$, $f_i = \begin{pmatrix} y_i - \mu_i \\ s_i - \sigma_i \end{pmatrix}$, $s_i^T = (s_{i11}, s_{i12}, \cdots, s_{idd})$ with $s_{ikl} = s_{ikl}(\beta) = (y_{ik} - \mu_{ik})(y_{il} - \mu_{il})$ and $\sigma_i^T = (\sigma_{i11}, \sigma_{i12}, \cdots, \sigma_{i22}, \cdots, \sigma_{idd})$. Compared to GEE2, GEE1 (the GEE by Liang and Zeger, 1986) can be inefficient for the estimation of the correlation parameter ρ. The GEE2 approach should be applied if both the mean and covariance parameters are of interest. However, GEE2 is not robust to misspecification of correlation structure. Another problem for GEE2 is that it can become computationally infeasible as the observation times (or cluster size) n_i gets large since there are $\binom{n_i}{2}$ estimating equations for the correlation parameters (Carey, Zeger and Diggle, 1993). Therefore, GEE1 should be used when the correlation parameter ρ is considered as a nuisance parameter.

When the working correlation structure is misspecified, one pitfall of the GEE approach is that in some cases $\hat{\rho}$ does not exist or does not converge which can lead to a breakdown of the asymptotic properties of the regression parameters (Crowder 1995). Crowder (1995) suggested two approaches to avoid the problem. One suggestion is to use only estimating equations which have a guaranteed solution. Another suggestion is to minimize some objective function with respect to ρ.

Adopting the idea of Crowder (1995), Chaganty (1997) presented a new method called quasi-least square (QLS) for estimating the correlation parameters. By the principle of generalized least squares, which requires minimizing the quadratic form

$$\begin{aligned} Q_\phi(\beta,\rho) &= \frac{1}{\phi}\sum_{i=1}^{N}(y_i-\mu_i(\beta))^T W_i^{-1}(y_i-\mu_i(\beta)) \\ &= \frac{1}{\phi}\sum_{i=1}^{N}(y_i-\mu_i(\beta))^T A_i^{-1/2}(\beta)R^{-1}(\rho)A_i^{-1/2}(\beta)(y_i-\mu_i(\beta)), \end{aligned} \tag{2.4.4}$$

estimating equations for ρ are obtained by taking the partial derivative of $Q_\phi(\beta,\rho)$ with respect to ρ and equating it to zero. The resulting estimating equations are

$$\sum_{i=1}^{N} Z_i^T \frac{\partial R^{-1}(\rho)}{\partial \rho_j} Z_i = 0, \quad 1 \le j \le q, \tag{2.4.5}$$

where $Z_i = A_i^{-1/2}(\beta)(y_i-\mu_i(\beta)), 1 \le i \le N$. And the set of estimating equations for β is exactly the GEEs proposed by Liang and Zeger (1986). Solutions for ρ and β are obtained by an iteration procedure. Shults and Chaganty (1998) then applied this QLS method to the analysis of serially correlated data.

The QLS estimates of the regression parameters β and the dispersion parameter ϕ are consistent even if the working correlation structure is misspecified. The estimates of the correlation parameters, however, are asymptotically biased. Chaganty and Shults (1999) proposed a modified (C-QLS) estimate of the correlation parameter to eliminate the asymptotic bias for the following working correlation structures: the unstructured matrix, the exchangeable, tridiagonal, and autoregressive structures.

Another method to bypass the pitfall is to use quadratic inference functions (QIF) developed by Qu, Lindsay and Li (2000). This method is based on the fact that the inverse of the working correlation matrix can be expressed by the linear combination of known basis matrices $M_1, \cdots, M_m$. That is,

$$R^{-1}(\rho) = \sum_{l=1}^{m} a_l M_l, \tag{2.4.6}$$

where $a_1, \cdots, a_m$ are unknown constants. Plugging this expression into the GEE (2.4.1), we have

$$\sum_{i=1}^{N} \frac{\partial \mu_i^T}{\partial \beta} A_i^{-\frac{1}{2}} (a_1 M_1 + \ldots + a_m M_m) A_i^{-\frac{1}{2}} (y_i - \mu_i) = 0.$$

Define the extended score g_N as

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

with $\dot{\mu}_i = \frac{\partial \mu_i}{\partial \beta}$. It is not possible to solve $g_N(\beta) = 0$ since the vector $g_N(\beta)$ contains more estimating equations than parameters. By the theory of generalized method of moments (Hansen, 1982), the estimate of β is obtained by minimizing the quadratic inference function $Q_N(\beta)$, that is,

$$\hat{\beta} = \arg\min_{\beta} Q_N(\beta),$$

where the QIF $Q_N(\beta)$ is defined to be

$$Q_N(\beta) = g_N^T C_N^{-1} g_N \tag{2.4.7}$$

and $C_N = \sum_{i=1}^{N} g_i(\beta) g_i(\beta)^T$. The QIF method avoids estimating the correlation parameter ρ in GEE. Qu, Lindsay and Li (2000) showed by simulations that if the working correlation structure is misspecified, the QIF approach results in more efficient regression estimates compared with the GEE method. On the other hand, when the working structure is correctly specified, the two methods produce equally efficient estimates.

To avoid misspecification of the working correlation structure, Ye and Pan (2006) proposed an approach for joint modelling of the mean and the covariance structures of longitudinal data within the framework of generalized estimating equations. They used the modified Cholesky decomposition to decompose the within-subject covariance matrices and then model the within-subject correlation and variation by simple regression models. The modified Cholesky decomposition of the within-subject covariance matrices Σ_i is given by $T_i' \Sigma_i T_i = D_i$, where T_i is a unique lower triangular matrix with 1's on the diagonal and D_i is a unique diagonal matrix. The Cholesky decomposition has an statistical interpretation

$$\hat{y}_{ij} = \mu_{ij} + \sum_{k=1}^{j-1} \phi_{ijk}(y_{ik} - \mu_{ik}),$$

where the negatives of the autoregressive coefficients ϕ_{ijk} are given by the below diagonal entries of T_i. Furthermore, $D_i = \mathrm{diag}\{\sigma_{ij}^2\}$ such that $\sigma_{ij}^2 = \mathrm{Var}(\varepsilon_{ij})$, where $\varepsilon_{ij} = y_{ij} - \hat{y}_{ij}$.

The joint modelling of the mean and covariance structures is based on three generalized linear models for the mean μ_{ij}, generalized autoregressive parameters ϕ_{ijk}

and the prediction error variances σ_{ij}^2:

$$g(\mu_{ij}) = x_{ij}^T\beta,\ \phi_{ijk}^T = t_{ijk}^T\gamma,\ \log\sigma_{ij}^2 = z_{ij}^T\lambda, \tag{2.4.8}$$

where x_{ij}, t_{ijk} and z_{ij} are column vectors of covariates, β, γ and λ are the associated parameters. The parameters β, γ and λ are estimated by jointly solving the corresponding generalized estimating equations in terms of the generalized linear models (2.4.8). See equation (4) in Ye and Pan (2006). The resulting estimators are shown to be consistent and asymptotically Normally distributed.

2.5. Gaussian copula regression models

Song (2000) developed a class of multivariate dispersion models generated from the multivariate Gaussian copula. These models enable us to analyze correlated (longitudinal) non-normal data in a way analogous to that of multivariate normal data.

The Gaussian copula model is described as what follows. Let $y = (y_1, \dots, y_m)$ be a vector of correlated variables and suppose each y_i is from a dispersion model (DM) of Jørgenson (1997) with density

$$f(y_i; \mu_i, \sigma_i^2) = a(y_i; \sigma_i^2)\exp\left\{-\frac{1}{2\sigma_i^2}d(y_i;\mu_i)\right\},$$

where d is the regular unit deviance. The exponential dispersion (ED) family or the exponential family with density (2.2.1) with $a(\phi) = \phi$, denoted by $ED(\mu, \phi)$, is a special class of dispersion models.

Denote the marginal cumulative distribution function (CDF) of y_j by $G_j(y_j)$ or $G_j(y_j; \mu_j, \phi_j)$. Then a joint CDF with m ED margins constructed by the Gaussian copula is given by

$$F(y; \mu, \phi, \Gamma) = C\{G_1(y_1; \mu_1, \phi_1), \dots, G_m(y_m; \mu_m, \phi_m)|\Gamma\}, \tag{2.5.1}$$

where $C(\cdot)$ is the m-variate Gaussian copula with the CDF given by

$$C(u|\Gamma) = \Phi_m\{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_m)|\Gamma\},$$

$$u = (u_1, \ldots, u_m)^T \in (0,1)^m.$$

In the above, Φ_m and Φ are the CDFs of m-variate normal $N_m(0, \Gamma)$ with a correlation matrix Γ and the standard univariate normal $N(0,1)$ margins. The resulting distribution with CDF (2.5.1) is called MED (multivariate ED) family. The (i,j)th element of the correlation matrix Γ is given by

$$\gamma_{ij} = \mathrm{Corr}[\Phi^{-1}\{G_i(y_i)\}, \Phi^{-1}\{G_j(y_j)\}]. \tag{2.5.2}$$

Using a third-order approximation (Jørgenson, 1997) to marginal normal scores on the basis of the deviance residual $r = r(y) = \pm d^{1/2}(y;\mu)$, Song (2000) approximated the density of the model by

$$g(y) = |\Gamma|^{-1/2} \prod_{i=1}^{m} a(y_i; \sigma_i^2) \exp\left\{-\frac{1}{2} r^T(y;\mu)\Sigma^{-1} r(y;\mu)\right\}, \tag{2.5.3}$$

where $\Sigma = \mathrm{diag}(\sigma_i)\Gamma\,\mathrm{diag}(\sigma_i)$. It is noted that (2.5.3) is of the form of the density of a multivariate normal distribution.

In longitudinal data analysis the vector outcomes might be of mixed types. For example, the responses contain continuous variables and binary response variables. The traditional method is to separate the responses and fit the two marginal models separately. This method might result in efficiency loss because the correlations between the two types of response variables are ignored. Song, Li and Yuan (2009) applied the Gaussian copula method to jointly analyze the regression model of continuous, discrete and mixed correlated outcomes. This model is a multivariate analogue of the univariate GLM and as claimed by them it provides an efficiency gain in the estimation of the regression parameters.

CHAPTER 3

Gaussian Estimation for Longitudinal Binary Data

3.1. Introduction

Correlated binary response data arise in many longitudinal studies in which the main purpose is to study the effects of the covariates on the correlated binary responses. For example, in the Six Cities study of the health effects of air pollution, analyzed by Fitzmaurice and Laird (1993), one of the purposes is to determine whether maternal smoking significantly affects the wheezing status of children.

One method of analyzing binary longitudinal response data is by the method of generalized estimating equations (GEE) proposed by Liang and Zeger (1986) in which a working correlation matrix for the responses for each individual is used (see, for example, Prentice, 1988 and Fitzmaurice, Laird and Rotnitzky, 1993). However, there is a history of controversy over choosing the working correlation structure $R(\rho)$ in GEE. For example, Crowder (1995) found that in some cases the parameters involved in the working correlation matrix are subject to an uncertainty of definition which can lead to a breakdown of asymptotic properties of the estimators (see also Crowder, 2001). Further, the misspecification of the correlation structure can result in loss of efficiency of the regression parameters (Wang and Carey, 2003).

Likelihood based methods are also available. For example, Lipsitz, Fitzmaurice, Sleeper and Zhao (1995) used a likelihood for the binary responses based on the Bahadur representation and Fitzmaurice and Laird (1993) used an exponential likelihood based on odds ratios. The likelihood approaches are rather complicated except in some special cases, such as, the analysis of paired binary data (Prentice, 1988).

Stefanescu and Turnbull (2005) used the likelihood approach based on a multivariate probit (MP) model for the analysis of longitudinal binary response data. Chaganty and Joe (2004) showed that the GEE method with the working correlation matrix $R(\rho)$ has good efficiency relative to the likelihood approach using a MP model. However, they recommended that $R(\rho)$ should be a weight matrix rather than a correlation matrix of binary responses and they suggest a method of choosing this weight matrix.

Whittle (1961) introduced the Gaussian estimation procedure in time series which uses the normal log-likelihood, without assuming that the data are normally distributed. The purpose of this chapter is to develop and investigate a Gaussian estimation procedure for the estimation of regression parameters in correlated (longitudinal) binary response data and compare this method with the GEE method and the weighted GEE method of Chaganty and Joe (2004). The motivation of this comes from the good properties of the Gaussian estimation procedure in other applications. For example, Crowder (1985) showed by simulation that a Gaussian estimate of the correlation parameter of equi-correlated clustered binary data has high efficiency. Paul and Islam (1998), again, by simulation, showed that a Gaussian estimator of the overdispersion parameter in clustered binomial data has best efficiency in comparison to likelihood, quasi-likelihood and extended quasi-likelihood estimates. Wang and Zhao (2007) used Gaussian estimation for the analysis of longitudinal data when the covariance function is modelled by additional variance parameters to the mean parameters. The variance parameters are estimated by Gaussian estimation and the regression parameters are estimated by the GEE method (see Wang and Zhao, 2007 for more details). See also Hand and Crowder (1996) for more applications.

In this chapter, we use the Gaussian log-likelihood function as an estimating function for the regression parameters. This is different from the method of Wang and Zhao (2007) in which the regression parameters are estimated by the GEE method. As in the GEE we use a working correlation matrix for the responses of each individual. Consistency of the parameter estimates is ensured by a carefully chosen robust working correlation matrix. A Newton-Raphson algorithm is derived for estimating the regression parameters from the Gaussian likelihood estimating equations for known correlation parameters. The correlation parameters of the working correlation matrix are estimated by the method of moments. A two-step iterative procedure is suggested for the joint estimation of the regression parameters and the correlation parameters. We show that the estimates of the regression parameters and the correlation parameters are consistent if the working correlation matrix considered is unstructured irrespective of whether the true correlation structure is unstructured, general autocorrelation, AR(1) or exchangeable. Similarly, the estimates of the regression parameters and the correlation parameters of the working correlation matrix are consistent when the working correlation matrix considered is general autocorrelation irrespective of whether the true correlation structure is general autocorrelation, AR(1) or exchangeable. Asymptotic variances of the Gaussian estimates of the regression parameters are also obtained. As far as we can find, these results for the Gaussian estimation procedure for correlated binary response data are new.

A simulation study was conducted to compare efficiency properties of twelve estimators of the regression parameters, namely, the maximum likelihood estimates using a multivariate probit (MP) model, four versions of the Gaussian estimates developed here, five versions of the generalized estimating equations (GEE) and two versions from a recent weighted GEE by Chaganty and Joe (2004). Efficiency results are obtained for all the methods using four different data sets generated from the MP model with latent correlation structures (i) exchangeable, (ii) AR(1), (iii) general autocorrelation and (iv) unstructured.

The Gaussian estimation procedure is developed and the theoretical results are obtained in Section 3.2. The Simulation study is conducted in Section 3.3. Two data sets are analyzed in Section 3.4. and a discussion follows in Section 3.5.

3.2. Gaussian Estimation of the Regression Parameters

3.2.1. Estimation of the regression parameters.

For simplicity, assume the number of observations of each subject has a common value d. Let $y_i = (y_{i1}, \dots, y_{id})^T$ be the $d \times 1$ vector of binary responses with a $d \times p$ design matrix $X_i = (x_{i1}, \dots, x_{id})^T$ for the ith subject, $i = 1, \dots, N$. Assume that the N subjects are independent while the repeated measurements y_{ij} taken on each subject are correlated. Define $\mu_i = \mathrm{E}(y_i|X_i) = (\mu_{i1}, \dots, \mu_{id})^T$ to be the expectation of y_i conditional on X_i and suppose $\mu_i = F(X_i\beta)$, where β is a $p \times 1$ vector of regression parameters of interest and F^{-1} is the link function. For the binary response data we consider the logit and probit link functions. The variance of y_{ij} is given by $v(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$.

Let $R(\rho)$ be a $d \times d$ working correlation matrix completely specified by the parameter vector ρ of length q and $W_i = A_i^{1/2} R(\rho) A_i^{1/2}$ be the corresponding $d \times d$ working covariance matrix, where $A_i(\beta) = \mathrm{diag}\{\mu_{ij}(1 - \mu_{ij})\}, j = 1, \dots, d$, $i = 1, \dots, N$. Further, let I_d be an identity matrix of dimension d.

Then, the Gaussian log-likelihood is given by

$$l(\beta, \rho) = \sum_{i=1}^{N} l_i = -\frac{1}{2} \sum_{i=1}^{N} \left\{ \log\det[2\pi W_i] + (y_i - \mu_i)^T W_i^{-1} (y_i - \mu_i) \right\}. \qquad (3.2.1)$$

The Gaussian score function for the parameter β_k, $k = 1, \dots, p$, is given by

$$\frac{\partial l}{\partial \beta_k} = \sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta_k} \right)^T W_i^{-1}(y_i - \mu_i) + \frac{1}{2} \operatorname{tr} \left[\sum_{i=1}^{N} \left\{ W_i^{-1}(y_i - \mu_i)(y_i - \mu_i)^T - I_d \right\} W_i^{-1} \frac{\partial W_i}{\partial \beta_k} \right]. \tag{3.2.2}$$

See also Crowder (2001, p. 56). Note that the elements of the $d \times 1$ vector $\frac{\partial \mu_i}{\partial \beta_k}$ depend on the link function. Then, for given values of ρ, the maximum Gaussian likelihood estimates of the regression parameters are obtained by solving the system of p estimating equations

$$\frac{\partial l}{\partial \beta_k} = 0, \; k = 1, \dots, p, \tag{3.2.3}$$

simultaneously. To solve equations (3.2.3), we use the Newton-Raphson method. Now, let $\frac{\partial l}{\partial \beta} = (\frac{\partial l}{\partial \beta_1}, \dots, \frac{\partial l}{\partial \beta_p})^T$. Further, let $\frac{\partial^2 l}{\partial \beta \partial \beta^T} = \left\{ \frac{\partial^2 l}{\partial \beta_k \partial \beta_{k'}} \right\}$ be the corresponding $p \times p$ second derivative matrix. Explicit expressions for $\frac{\partial^2 l}{\partial \beta_k \partial \beta_{k'}}$ are given in Appendix A. Then, based on the Newton-Raphson method, the Gaussian estimates are updated according to

$$\hat{\beta}^{(s+1)} = \hat{\beta}^{(s)} + \left[\frac{\partial^2 l}{\partial \beta \partial \beta^T} \right]^{-1}_{\hat{\beta}^{(s)}} \frac{\partial l}{\partial \beta} \Big|_{\hat{\beta}^{(s)}}, \quad s = 1, 2, \dots. \tag{3.2.4}$$

Note that the Newton-Raphson procedure given above for estimating the regression parameter β is based on the assumption that the correlation parameters involved in the covariance matrix $W_i = A_i^{1/2}(\beta) R(\rho) A_i^{1/2}(\beta)$ are known. In what follows, we consider four popular working correlation matrices $R(\rho)$ (Liang and Zeger, 1986). Then following Sutradhar and Das (1999), Sutradhar (2003) and Wang and Carey (2003) we propose to estimate the correlation parameters of the working correlation matrices by the method of moments. The four working correlation structures considered here are:

i) exchangeable correlation structure in which the diagonal elements of $R(\rho)$ are 1 and the off-diagonal elements are ρ,

ii) AR(1) correlation structure in which the diagonal elements of $R(\rho)$ are 1 and the off-diagonal elements are $\rho^{|i-j|}$, $i \neq j$,

iii) the general autocorrelation structure

$$R(\rho_1,\dots,\rho_{d-1}) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{d-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{d-2} \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \rho_{d-1} & \rho_{d-2} & \rho_{d-3} & \cdots & 1 \end{bmatrix},$$

and iv) the unstructured correlation matrix (Liang and Zeger, 1986)

$$R = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1,d-1} \\ \rho_{12} & 1 & \rho_{23} & \cdots & \rho_{2,d-2} \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \rho_{1,d-1} & \rho_{2,d-2} & \rho_{3,d-3} & \cdots & 1 \end{bmatrix}. \tag{3.2.5}$$

Let $y^*_{ij} = (y_{ij} - \hat{\mu}_{ij})/\sqrt{\hat{\mu}_{ij}(1-\hat{\mu}_{ij})}$. Then, the method of moments estimate of (i) the common correlation coefficient ρ in the exchangeable correlation structure is

$$\hat{\rho} = \frac{\sum_{i=1}^N \sum_{j\neq k} y^*_{ij} y^*_{ik}}{(d-1)\sum_{i=1}^N \sum_{j=1}^d {y^*_{ij}}^2},$$

(ii) the common correlation coefficient ρ in the AR(1) correlation structure is

$$\hat{\rho} = \frac{\sum_{i=1}^N \sum_{j=2}^d y^*_{ij} y^*_{i,j-1}}{\sum_{i=1}^N \{\sum_{j=2}^{d-1} {y^*_{ij}}^2 + ({y^*_{i1}}^2 + {y^*_{id}}^2)/2\}},$$

(iii) the correlation parameter ρ_l in $R(\rho_1,\dots,\rho_{d-1})$ is

$$\hat{\rho}_l = \frac{\sum_{i=1}^N \sum_{j=1}^{d-l} y^*_{ij} y^*_{i,j+l}/(d-l)}{\sum_{i=1}^N \sum_{j=1}^d {y^*_{ij}}^2/d}, \quad l = 1,\dots,d-1.$$

(iv) Finally, the estimate of the unstructured correlation matrix is given by

$$\hat{R} = \sum_{i=1}^N \hat{A}_i^{-1/2} S_i S_i^T \hat{A}_i^{-1/2}/N, \text{ where } S_i = y_i - \hat{\mu}_i, i = 1,\dots,N,$$

and the diagonal elements are forced to be equal to 1.

The Newton-Raphson iterative procedure for estimating the regression parameters and the method of moments estimates of the correlation parameters are combined in a two-step iterative procedure which is described in what follows.

> Step 1: For given initial values β^0 of β and ρ^0, where ρ is the vector of correlation parameters (depending on the structure of the working correlation matrix chosen), estimate β via the formula (3.2.4). Denote this by β^1.
>
> Step 2: Obtain the elements of ρ by the method of moments described above using β^1. Denote this estimate of ρ by ρ^1.

Iterate between step 1 and step 2 until convergence.

3.2.2. Consistency of the estimates of the parameters.

We show in Appendix B that if the estimate $R(\hat{\rho})$ of the working correlation $R(\rho)$ converges to the true correlation matrix $C(\rho)$ in probability, then the estimating equations (3.2.3) are asymptotically, as $N \to \infty$, unbiased and therefore the estimator $\hat{\beta}$ obtained by solving the system of equations given by (3.2.3) is consistent. It then remains to show that $R(\hat{\rho})$ is consistent.

THEOREM 3.2.1. *The moment estimates of the correlation parameters of the unstructured correlation matrix are consistent whatever is the true correlation structure: unstructured, general autocorrelation, AR(1) or exchangeable.*

The proof of the theorem is given in Appendix C. However, the reverse is not true. For example, the moment estimate of the correlation parameter ρ of the exchangeable correlation structure is not consistent when the true correlation structure is any of the other three. Similarly, the moment estimates of the correlation parameters of the general autocorrelation structure are consistent when the true correlation structure is general autocorrelation, AR(1) or exchangeable. In this sense the unstructured correlation matrix is most robust against misspecification by other correlation structures. The next robust, of course, is the general autocorrelation structure.

Now, there is some circularity in the proofs, in that consistency of $\hat{\beta}$ requires consistency of $\hat{\rho}$ and vice versa. However, this problem can be overcome by using consistent estimate of, for example, β at the initial stage of the iterative procedure. That is, overall consistency of $\hat{\beta}$ and $\hat{\rho}$ are obtained if consistent initial estimates of β, such as the GEEs using independence working correlation structure, are used at step 1 of the two step iterative procedure described at the end of section 3.2.1.

3.2.3. Variance of $\hat{\beta}$.

In Appendix B we have shown that the estimating equations (3.2.3) are asymptotically, as $N \to \infty$, unbiased. So, by the general theory of unbiased estimating functions (Crowder, 1986 and Liang and Zeger, 1995), the estimator $\hat{\beta}$ by (3.2.3) is consistent and has asymptotic multivariate normal distribution $MVN(\beta, V_\beta)$, where V_β is given by

$$V_\beta = D^{-1}V(D^{-1})^T, \tag{3.2.6}$$

where $D = -\sum_{i=1}^{N} \mathrm{E}\left\{\frac{\partial^2 l_i}{\partial\beta_k \partial\beta_{k'}}\right\}$ and V is a $p \times p$ matrix with diagonal elements $\sum_{i=1}^{N} \mathrm{Var}(\frac{\partial l_i}{\partial \beta_k})$ and the (k, k')th off diagonal elements $\sum_{i=1}^{N} \mathrm{Cov}(\frac{\partial l_i}{\partial\beta_k}, \frac{\partial l_i}{\partial\beta_{k'}})$. Expressions for $\mathrm{E}\left\{\frac{\partial^2 l_i}{\partial\beta_k \partial\beta_{k'}}\right\}$, $\mathrm{Var}(\frac{\partial l_i}{\partial\beta_k})$ and $\mathrm{Cov}(\frac{\partial l_i}{\partial\beta_k}, \frac{\partial l_i}{\partial\beta_{k'}})$ are given in Appendix D.

In (3.2.6), the true covariance matrix Σ_i is estimated by $\widehat{\Sigma}_i = \hat{A}_i^{1/2}\hat{R}\hat{A}_i^{1/2}$. The variance V_β of $\hat{\beta}$ is estimated by replacing β and Σ_i with their estimates $\hat{\beta}$ and $\widehat{\Sigma}_i$ respectively.

3.3. Simulations

In this section we compare, by simulations, twelve estimators of the regression parameters, namely, the maximum likelihood estimates using a MP model, four versions of the Gaussian estimates, five versions of the GEE and two versions of the weighted GEE.

Following Chaganty and Joe (2004), we use the multivariate probit (MP) model as a data generation mechanism. The MP model is a commonly used model for multivariate binary data. It assumes that the binary response is the indicator of the event that an unobserved latent variable exceeds a given threshold. Let $y_i = (y_{i1}, \ldots, y_{id})^T$ be the d-dimensional vector of binary responses on the ith subject, $i = 1, \ldots, N$. Let $x_i = (x_{i1}, \ldots, x_{id})^T$ be a $d \times p$ covariate matrix. Let $Z_i = (z_{i1}, \ldots, z_{id})^T$ be a d-dimensional vector of latent variables such that $Z_i = x_i\beta + \epsilon_i,\ i = 1, \ldots, N$. The latent variable Z_i is assumed to follow a multivariate normal distribution with mean $x_i\beta$ and covariance $\Omega(\gamma)$, where γ is the latent correlation. The relationship between z_{ij} and y_{ij} in the MP model is given by

$$y_{ij} = \begin{cases} 1, & \text{if } z_{ij} > 0; \\ 0, & \text{otherwise.} \end{cases} \qquad j = 1, \ldots, d.$$

Thus $P(y_{ij} = 1|X_i) = P(z_{ij} > 0) = \Phi(\beta' x_{ij})$, where Φ is the standard normal distribution function. It can be seen that the correlation between any two binary responses y_{ij} and y_{ik} is given by

$$\mathrm{Corr}(y_{ij}, y_{ik}) = \frac{\Phi_2(v_j, v_k; \gamma) - \Phi(v_j)\Phi(v_k)}{[\Phi(v_j)\{1 - \Phi(v_j)\}\Phi(v_k)\{1 - \Phi(v_k)\}]^{1/2}},$$

where $\Phi_2(\omega_1, \omega_2; \gamma)$ is the bivariate normal distribution function with correlation γ, $v_j = \beta' x_{ij}$ and $v_k = \beta' x_{ik}$.

For simulating data from the MP model, we use the latent covariance matrix $\Omega(\gamma)$. For example, for generating binary data with exchangeable $R(\rho)$, we use the exchangeable correlation matrix $\Omega(\gamma)$. Note that the correlation ρ of the binary variables is always less than the latent correlation γ as shown in Chaganty and Joe (2004). Efficiencies of the estimates of the regression parameters are compared for all the methods using four different data sets generated from the MP model with latent correlation structures (i) exchangeable, (ii) AR(1), (iii) general autocorrelation and (iv) unstructured.

For the exchangeable or AR(1) model, we choose $\gamma = 0.5$. For general autocorrelation structure we use A for $\Omega(\gamma)$, where

$$A = \begin{pmatrix} 1.0 & 0.5 & 0.4 & 0.3 & 0.2 \\ 0.5 & 1.0 & 0.5 & 0.4 & 0.3 \\ 0.4 & 0.5 & 1.0 & 0.5 & 0.4 \\ 0.3 & 0.4 & 0.5 & 1.0 & 0.5 \\ 0.2 & 0.3 & 0.4 & 0.5 & 1.0 \end{pmatrix}.$$

For unstructured correlation, we use the following positive definite correlation matrix

$$U = \begin{pmatrix} 1.00 & 0.12 & 0.52 & 0.06 & 0.38 \\ 0.12 & 1.00 & 0.63 & 0.16 & 0.78 \\ 0.52 & 0.63 & 1.00 & 0.10 & 0.90 \\ 0.06 & 0.16 & 0.10 & 1.00 & 0.15 \\ 0.38 & 0.78 & 0.90 & 0.15 & 1.00 \end{pmatrix}.$$

For all correlation structures we choose probit link, $d = 5$, $p = 2$, $\mathbf{x}_{ij} = (1, x_{ij})^T$, where x_{ij} are taken as uniform random variables in the interval $[-1.0, 1.0]$, $\beta = (0.0, 0.5)$ and $N = 50, 80, 150$.

TABLE 3.1. $N\times$ average estimated variance for $\hat{\beta}_0$ and $\hat{\beta}_1$ by Gaussian estimation procedure using the four working correlation structures: data generated from MP model with latent (i) exchangeable $R(0.5)$; (ii) AR(1) $R(0.5)$; (iii) general autocorrelation matrix A and (iv) unstructured covariance matrix U; x_{ij} ~ uniform(-1,1); $p = 2$, $\beta_0 = 0.0, \beta_1 = 0.5$; observation times $d = 5$; based on 500 iterations.

N	Method	(i) $N \times \widehat{\text{Var}}(\hat{\beta})$	(ii) $N \times \widehat{\text{Var}}(\hat{\beta})$	(iii) $N \times \widehat{\text{Var}}(\hat{\beta})$	(iv) $N \times \widehat{\text{Var}}(\hat{\beta})$
	Gaussian-Exch	(0.783, 0.701)	(0.621, 0.850)	(0.705, 0.780)	(0.831, 0.837)
50	Gaussian-AR(1)	(0.419, 0.853)	(0.476, 0.831)	(0.512, 0.765)	(0.383, 1.030)
	Gaussian-Autocorr	(0.770, 0.661)	(0.587, 0.722)	(0.681, 0.710)	(0.810, 0.724)
	Gaussian-Unstr	(0.721, 0.591)	(0.552, 0.650)	(0.640, 0.636)	(0.602, 0.473)
	Gaussian-Exch	(0.775, 0.700)	(0.616, 0.855)	(0.691, 0.793)	(0.823, 0.845)
80	Gaussian-AR(1)	(0.408, 0.866)	(0.466, 0.835)	(0.499, 0.774)	(0.371, 1.052)
	Gaussian-Autocorr	(0.768, 0.675)	(0.584, 0.736)	(0.670, 0.730)	(0.804, 0.745)
	Gaussian-Unstr	(0.740, 0.633)	(0.561, 0.689)	(0.643, 0.681)	(0.610, 0.491)
	Gaussian-Exch	(0.766, 0.710)	(0.608, 0.857)	(0.681, 0.793)	(0.809, 0.855)
150	Gaussian-AR(1)	(0.396, 0.883)	(0.455, 0.845)	(0.491, 0.774)	(0.359, 1.070)
	Gaussian-Autocorr	(0.762, 0.696)	(0.581, 0.754)	(0.664, 0.744)	(0.793, 0.759)
	Gaussian-Unstr	(0.747, 0.672)	(0.570, 0.729)	(0.651, 0.718)	(0.617, 0.518)

For each N, we simulated 500 samples and obtained the estimates of β_0 and β_1 for each sample. We then calculated $N\times$ average estimated variance $(\sum_{i=1}^{500} \widehat{\text{Var}}(\hat{\theta}_i)/500)$, where $\hat{\theta}_i$ is either $\hat{\beta}_0$ or $\hat{\beta}_1$ for the ith sample.

We first compare the Gaussian estimation procedures with the four correlation structures discussed earlier, namely the exchangeable, the AR(1), the general autocorrelation and the unstructured correlation. The results are given in Table 3.1.

We see from the results in Table 3.1 that the Gaussian estimation procedure using AR(1) correlation structure produces the smallest variance estimates for β_0 and largest variance estimates for β_1 irrespective of the data generation procedure. Among the three other methods, in general, Gaussian estimation using exchangeable correlation structure produces largest variance estimates for both β_0 and β_1. The other two estimation procedures, in general, produce smallest estimated variance, although the

Gaussian estimation procedure using the unstructured correlation produces smallest variance estimates among these three methods.

We now compare the two Gaussian estimation procedures using the general autocorrelation structure (Gaussian-Autocorr) and the unstructured correlation matrix (Gaussian-Unstr) with the maximum likelihood (ML) estimates based on the MP model and GEE approaches. We consider GEE-independence (GEE-I), GEE-AR(1), GEE-exchangeable (GEE-ex), GEE-general autocorrelation (GEE-Autocorr), GEE-unstructured (GEE-un) and weighted GEE-exchangeable by Chaganty and Joe (GEE-CJ).

For all data sets, ML estimates were obtained using the exchangeable correlation structure. Results are similar for AR(1) correlation structure. For the estimation using Chaganty and Joe's method, we use the exchangeable correlation structure and the AR(1) correlation structure, both with $\rho = 0.3$ (following their guidelines for choosing ρ). Thus, we use two versions of GEE-CJ, henceforth named as GEE-CJ(EX) and GEE-CJ(AR(1)). Note that data were generated using latent correlation $\gamma = 0.5$. According to the recommendations of Chaganty and Joe (2004), the value of ρ to be taken for the estimation of the regression parameters should be less than 0.5. We examined efficiency results of the above two methods using other values of ρ which satisfy this requirement, such as, $\rho = 0.2$ and the results are found to be similar. Results of $N\times$ average estimated variance for $\hat{\beta}_0$ and $\hat{\beta}_1$ are given in Table 3.2.

Results in Table 3.2 show that the performance of GEE-I is the worst, at least in terms of $\hat{\beta}_1$, producing the largest variance irrespective of the data generation mechanism. Again, irrespective of the data generation mechanism, in terms of $\hat{\beta}_1$, Gaussian-Unstr performs the best, producing the smallest variance and the next best

Table 3.2. $N\times$ average estimated variance for $\hat{\beta}_0$ and $\hat{\beta}_1$ by ML, Gaussian-Autocorr, Gaussian-Unstr and GEE methods: data generated from MP model with latent (i) exchangeable $R(0.5)$; (ii) AR(1) $R(0.5)$; (iii) general autocorrelation matrix A and (iv) unstructured covariance matrix U; x_{ij} ~ uniform(-1,1); p = 2, $\beta_0 = 0.0, \beta_1 = 0.5$; observation times d = 5; based on 500 iterations.

N	Method	(i) $N \times \widehat{\text{Var}}(\hat{\beta})$	(ii) $N \times \widehat{\text{Var}}(\hat{\beta})$	(iii) $N \times \widehat{\text{Var}}(\hat{\beta})$	(iv) $N \times \widehat{\text{Var}}(\hat{\beta})$
	ML	(0.720, 0.836)	(0.552, 0.883)	(0.582, 0.854)	(0.634, 0.888)
	Gaussian-Autocorr	(0.774, 0.650)	(0.581, 0.720)	(0.677, 0.714)	(0.811, 0.718)
	Gaussian-Unstr	(0.725, 0.586)	(0.554, 0.637)	(0.638, 0.643)	(0.616, 0.459)
	GEE-I	(0.740, 1.032)	(0.572, 1.013)	(0.654, 0.999)	(0.644, 1.023)
50	GEE-AR(1)	(0.749, 0.901)	(0.563, 0.869)	(0.651, 0.870)	(0.649, 0.986)
	GEE-ex	(0.742, 0.830)	(0.574, 0.923)	(0.657, 0.860)	(0.646, 0.887)
	GEE-Autocorr	(0.731, 0.801)	(0.558, 0.841)	(0.645, 0.821)	(0.644, 0.839)
	GEE-un	(0.713, 0.765)	(0.540, 0.799)	(0.628, 0.779)	(0.553, 0.686)
	GEE-CJ(EX)	(0.741, 0.832)	(0.575, 0.936)	(0.652, 0.884)	(0.646, 0.893)
	GEE-CJ(AR(1))	(0.748, 0.905)	(0.564, 0.874)	(0.646, 0.885)	(0.661, 1.043)
	ML	(0.722, 0.825)	(0.547, 0.876)	(0.570, 0.848)	(0.632, 0.875)
	Gaussian-Autocorr	(0.761, 0.685)	(0.586, 0.734)	(0.666, 0.734)	(0.803, 0.733)
	Gaussian-Unstr	(0.740, 0.634)	(0.561, 0.696)	(0.643, 0.688)	(0.613, 0.497)
	GEE-I	(0.731, 1.035)	(0.574, 1.003)	(0.649, 1.014)	(0.648, 1.026)
80	GEE-AR(1)	(0.738, 0.908)	(0.564, 0.862)	(0.645, 0.879)	(0.653, 0.995)
	GEE-ex	(0.733, 0.835)	(0.575, 0.913)	(0.650, 0.878)	(0.649, 0.893)
	GEE-Autocorr	(0.735, 0.808)	(0.560, 0.849)	(0.640, 0.841)	(0.644, 0.845)
	GEE-un	(0.715, 0.790)	(0.551, 0.816)	(0.630, 0.815)	(0.563, 0.709)
	GEE-CJ(EX)	(0.733, 0.836)	(0.575, 0.923)	(0.653, 0.882)	(0.650, 0.898)
	GEE-CJ(AR(1))	(0.738, 0.910)	(0.565, 0.864)	(0.646, 0.881)	(0.664, 1.055)
	ML	(0.730, 0.826)	(0.542, 0.866)	(0.565, 0.838)	(0.674, 0.869)
	Gaussian-Autocorr	(0.762, 0.699)	(0.581, 0.757)	(0.667, 0.743)	(0.794, 0.768)
	Gaussian-Unstr	(0.752, 0.669)	(0.570, 0.726)	(0.655, 0.718)	(0.615, 0.530)
	GEE-I	(0.738, 1.017)	(0.574, 1.005)	(0.652, 1.011)	(0.656, 1.032)
150	GEE-AR(1)	(0.746, 0.888)	(0.564, 0.865)	(0.648, 0.883)	(0.661, 1.001)
	GEE-ex	(0.738, 0.814)	(0.574, 0.914)	(0.652, 0.874)	(0.656, 0.893)
	GEE-Autocorr	(0.737, 0.819)	(0.564, 0.857)	(0.645, 0.846)	(0.646, 0.856)
	GEE-un	(0.730, 0.793)	(0.557, 0.840)	(0.639, 0.831)	(0.578, 0.730)
	GEE-CJ(EX)	(0.738, 0.815)	(0.574, 0.924)	(0.654, 0.871)	(0.656, 0.897)
	GEE-CJ(AR(1))	(0.745, 0.888)	(0.565, 0.866)	(0.649, 0.879)	(0.671, 1.053)

is Gaussian-Autocorr. Only when the data are generated using unstructured covariance matrix, does GEE-un have a slight edge over Gaussian-Autocorr. The estimate of the variance of $\hat{\beta}_0$ does not seem to differ much irrespective of the data generation mechanism and the method of estimation, although, Gaussian-Autocorr seems to produce a larger variance estimate when data are generated using the unstructured covariance matrix.

The simulation study was extended to compare bias and MSE. Again, based on 500 simulated samples, we obtained (a) average bias$(\hat{\theta}_i) = \sum_{i=1}^{500}(\hat{\theta}_i - \theta_i)/500$ and (b) $N\times$ average MSE ($\sum_{i=1}^{500}(\hat{\theta}_i - \theta_i)^2/500)$. To save space we only summarize the results (not given here) of the simulation.

Our simulations show that biases of the estimates by all procedures compared are small. In terms of the MSE, the performance of GEE-I is the worst in general. When data are simulated using unstructured correlation structure, Gaussian-Unstr is the best for the estimation of β_1, agreeing with the results shown in terms of estimated variances. For the estimation of β_0, no method seems to perform better than any other. For data with other correlation structures there do not seem to be any significant differences in efficiency both for the estimation of β_0 and β_1.

We conducted a further simulation study to compare these twelve estimators by generating correlated binary data with specified marginal means and correlations (Qaqish, 2003). Simulation results not reported here show similar conclusions.

3.4. Examples

Example 1: As a first example we consider a subset of data from the Six Cities study, a longitudinal study of the health effects of air pollution that was analyzed by Fitzmaurice and Laird (1993). The data set contains complete records on 537 children from Steubenville, Ohio, each of whom was examined annually at ages 7 through 10. The repeated binary response is the wheezing status (1=yes, 0=no) of a

TABLE 3.3. Results of the regression analysis of the wheezing status data; estimates of β_0, β_1, β_2 and β_3 of the model (3.4.1) with standard errors in parenthesis using maximum likelihood method based on the MP model, four Gaussian estimation methods and six GEE procedures; with probit link.

Method	Parameters			
	Intercept	Age	MS	Age*MS
ML-Exch	-1.1195(0.0611)	-0.0778(0.0307)	0.1610(0.1004)	0.0385(0.0500)
ML-AR(1)	-1.1296(0.0590)	-0.0819(0.0368)	0.1573(0.0972)	0.0438(0.0595)
Gaussian-Exch	-1.1255(0.0648)	-0.0829(0.0273)	0.1614(0.1056)	0.0391(0.0448)
Gaussian-AR(1)	-1.1562(0.0643)	-0.0839(0.0354)	0.1645(0.1034)	0.0408(0.0592)
Gaussian-Autocorr	-1.1252(0.0650)	-0.0846(0.0294)	0.1632(0.1059)	0.0410(0.0485)
Gaussian-Unstr	-1.1228(0.0649)	-0.0818(0.0289)	0.1598(0.1059)	0.0381(0.0477)
GEE-I	-1.1259(0.0634)	-0.0768(0.0313)	0.1709(0.1028)	0.0367(0.0486)
GEE-ex	-1.1258(0.0634)	-0.0768(0.0313)	0.1708(0.1028)	0.0367(0.0486)
GEE-AR(1)	-1.1359(0.0638)	-0.0800(0.0318)	0.1599(0.1035)	0.0426(0.0497)
GEE-Autocorr	-1.1289(0.0634)	-0.0780(0.0314)	0.1679(0.1028)	0.0390(0.0488)
GEE-un	-1.1299(0.0634)	-0.0771(0.0314)	0.1638(0.1030)	0.0354(0.0490)
GEE-CJ(EX)	-1.1258(0.0634)	-0.0768(0.0313)	0.1708(0.1028)	0.0367(0.0486)
GEE-CJ(AR(1))	-1.1331(0.0636)	-0.0792(0.0316)	0.1634(0.1031)	0.0413(0.0492)

child at each occasion. The purpose of the study is to model the probability of the wheezing status as a function of the child's age, his/her mother's maternal smoking habit (a binary variable MS with 1 if the mother smoked regularly and 0 otherwise) and their interactions. We consider the same marginal model used by Fitzmaurice and Laird (1993) with a probit link

$$\text{probit}(\mu) = \beta_0 + \beta_1\text{Age} + \beta_2\text{MS} + \beta_3\text{Age*MS}, \tag{3.4.1}$$

where 'age' is the age in years since the child's 9th birthday.

Estimates of the regression parameters of model (3.4.1) and their standard errors by all the methods discussed are given in Table 3.3. Estimates of the correlation parameters by all the methods are given in Table 3.5.

The standard errors of the estimates of β_0 and β_2 by the maximum likelihood method are the smallest and there does not appear to be a large difference among the standard errors of the estimates by all other methods. For β_1 and β_3, it appears that the estimates by the Gaussian estimation procedures, except Gaussian-AR(1), produce the smallest standard errors, providing some support that the estimates of the regression parameters by the Gaussian estimation procedure using the general autocorrelation structure and unstructured correlation have the highest efficiency.

Example 2: The second example uses a subset of the data from the Coronary Risk Factor Study by Woolson and Clarke (1984). The dataset contains records of 1014 children from Muscatine, Iowa, who were 7-9 years old in 1977. Height and weight were measured on each child in three survey years, 1977, 1979 and 1981. For each survey year, the median weight was calculated for each gender and 1 inch of height. Children with relative weight greater than 110% of the median weight in their respective stratum were classified as obese. The repeated binary response of interest is whether the child is described as being obese or not (1=yes, 0=no) at each occasion. Data on many children are incomplete, and only 460 children had complete data from all the three occasions. We analyze only the complete data (see Table 2 in Fitzmaurice, Laird and Lipsit, 1994). One of the objectives of this study was to determine the effects of age and gender on risk of obesity in children. Fitzmaurice, Laird and Lipsit (1994) analyzed these data using a logit link. Here we consider the marginal model with a probit link

$$\text{probit}(\mu) = \beta_0 + \beta_1\text{Age} + \beta_2\text{Gender} + \beta_3\text{Age*Gender}, \tag{3.4.2}$$

where Gender=1 if the child is female, 0 otherwise.

Estimates of the regression parameters of model (3.4.2) and their standard errors by the Gaussian and the GEE estimation procedures are given in Table 3.4. Estimates of the correlation parameters by all the methods are given in Table 3.5. The standard

TABLE 3.4. Results of the regression analysis of the complete Mluscatinie Study data; estimates of β_0, β_1, β_2 and β_3 of the model (3.4.2) with standard errors in parenthesis using four Gaussian estimation methods and six GEE procedures; with probit link.

Method	Parameters			
	Intercept	Age	Gender	Age*Gender
ML-Exch	-0.8760(0.2255)	0.0048(0.0212)	-0.8374(0.3436)	0.0866(0.0322)
ML-AR(1)	-0.8371(0.2692)	0.0014(0.0255)	-0.8140(0.3810)	0.0829(0.0358)
Gaussian-Exch	-0.9227(0.2221)	0.0098(0.0209)	-0.8391(0.3118)	0.0871(0.0295)
Gaussian-AR(1)	-0.9809(0.3228)	0.0109(0.0317)	-0.8347(0.4552)	0.0854(0.0449)
Gaussian-Autocorr	-0.9239(0.2238)	0.0099(0.0211)	-0.8382(0.3142)	0.0870(0.0297)
Gaussian-Unstr	-0.8896(0.2247)	0.0061(0.0210)	-0.8776(0.3148)	0.0919(0.0296)
GEE-I	-0.8701(0.2571)	0.0039(0.0239)	-0.8374(0.3526)	0.0882(0.0328)
GEE-ex	-0.8701(0.2571)	0.0039(0.0239)	-0.8360(0.3525)	0.0882(0.0328)
GEE-AR(1)	-0.8522(0.2518)	0.0039(0.0236)	-0.8716(0.3525)	0.0874(0.0328)
GEE-Autocorr	-0.8700(0.4492)	0.0039(0.0418)	-0.8240(0.6205)	0.0870(0.0571)
GEE-un	-0.8476(0.2546)	0.0017(0.0237)	-0.8896(0.3518)	0.0934(0.0327)
GEE-CJ(EX)	-0.8701(0.2571)	0.0039(0.0239)	-0.8361(0.3525)	0.0882(0.0328)
GEE-CJ(AR(1))	-0.8526(0.2520)	0.0039(0.0236)	-0.8709(0.3525)	0.0875(0.0328)

errors of the estimates of all regression parameters by Gaussian-Exch, Gaussian-Autocorr and Gaussian-Unstr are the smallest.

When Chaganty and Joe's method is used, we estimate the regression parameters using GEE-ex or GEE-AR(1) with $\rho = 0.3$ in example 1 and $\rho = 0.5$ in example 2. In both examples, the estimates of the working correlation parameters are always smaller than the maximum likelihood estimates of the latent correlation γ providing support for Chaganty and Joe's (Chaganty and Joe, 2004) finding.

TABLE 3.5. Estimates of the correlation parameters by different methods for the two examples.

Method	Example 1	Example 2
ML-Exch	0.60	0.76
ML-AR(1)	0.67	0.80
Gaussian-Exch	0.35	0.51
Gaussian-AR(1)	0.30	0.34
Gaussian-Autocorr	(0.40,0.31,0.30)	(0.51,0.50)
Gaussian-Unstr	$\begin{pmatrix} 1 & 0.35 & 0.31 & 0.30 \\ & 1 & 0.47 & 0.32 \\ & & 1 & 0.38 \\ & & & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.53 & 0.50 \\ & 1 & 0.48 \\ & & 1 \end{pmatrix}$
GEE-ex	0.35	0.51
GEE-AR(1)	0.40	0.51
GEE-Autocorr	(0.40, 0.31, 0.30)	(0.51,0.50)
GEE-un	$\begin{pmatrix} 1 & 0.35 & 0.31 & 0.30 \\ & 1 & 0.47 & 0.32 \\ & & 1 & 0.38 \\ & & & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.53 & 0.50 \\ & 1 & 0.48 \\ & & 1 \end{pmatrix}$
GEE-CJ(EX)	0.3	0.5
GEE-CJ(AR(1))	0.3	0.5

3.5. Discussion

In this chapter, we developed a Gaussian estimation procedure involving a working correlation matrix for the estimation of the regression parameters in longitudinal binary response data. It is interesting to see that the first part of the Gaussian score function (3.2.2) is the GEE function and the second part can be regarded as a correction term for obtaining estimates of regression parameters with higher efficiency. To preserve the (asymptotic) unbiasedness of the Gaussian estimating equations (3.2.3), the second part of (3.2.2) should be convergent to zero asymptotically. For this purpose, we prefer to choose working correlation matrices with robust structures.

We showed that the estimates of the regression parameters and the correlation parameters are consistent if the working correlation matrix considered is unstructured irrespective of whether the true correlation structure is unstructured, general autocorrelation, AR(1) or exchangeable. Thus, the Gaussian estimates of the regression parameters using an unstructured working correlation matrix do not suffer from the pitfalls that Crowder (1995, p. 408) discusses regarding the GEE estimates. Further, the estimates of the regression parameters and the correlation parameters of the working correlation matrix are consistent when the working correlation matrix considered is general autocorrelation irrespective of whether the true correlation structure is general autocorrelation, AR(1) or exchangeable. In this sense the unstructured correlation matrix is most robust against misspecification by other correlation structures. The next most robust is the general autocorrelation structure.

It shows in Section 2.5 that the Gaussian estimating functions can be obtained by an approximation to marginal normal scores on the basis of the deviance residual (see equation (2.5.3)) in the Gaussian copula regression model (Song, 2000). This provides another theoretical justification of the application of Gaussian estimation to longitudinal binary data analysis. Though both Gaussian estimation and Gaussian copula regression model use the density function (or cumulative distribution function) of a multivariate normal distribution, there are two major differences between these two methods. First, the covariance matrix in Gaussian estimation models the correlation between any two response variables directly while the covariance matrix in the Gaussian copula model measures the correlation between two normal scores, $\Phi^{-1}\{G_i(y_i)\}$ and $\Phi^{-1}\{G_j(y_j)\}$ (see equation (2.5.2)). Second, the joint density function of the response variables is unknown in Gaussian estimation. The Gaussian score function is for estimation purpose only. However, the density function of the response variables in the Gaussian copula regression model is given on the basis of the Gaussian copula. The comparison of Gaussian estimation and estimation using Gaussian copula regression models (Song, Li and Yuan, 2009) is of interest for a future research.

Twelve estimators of the regression parameters consisting of the maximum likelihood estimates based on the multivariate probit (MP) model, four Gaussian estimates, five GEE estimates and two weighted GEE estimates are compared by simulations. Efficiencies of the estimates of the regression parameters are compared for all the methods using four different data sets generated from the MP model with latent correlation structures (i) exchangeable, (ii) AR(1), (iii) general autocorrelation and (iv) unstructured. Simulations show that Gaussian estimates of the regression parameters, using the unstructured correlation matrix of the responses for a subject, are, in general, more efficient than those by the other eleven methods irrespective of the data generation method. This also shows evidence of the robustness of this method even if the correlation structure is not unstructured, but one of the other three: exchangeable, AR(1) and general autocorrelation.

We have written a SAS macro %Gaussian which can be implemented to estimate the regression parameters, the parameters of the working correlation matrix and the variances and standard errors of the estimates of the regression parameters.

The Gaussian estimation methodology developed here for binary data can be easily extended to other data distribution situations by changing the link and the variance functions. For example, for Poisson data y_{ij}, we need a log link implying that $\mu_i = \exp(X_i\beta)$ and variance function $v(\mu_{ij}) = \mu_{ij}$.

CHAPTER 4

Bias Correction for GEE Estimation

4.1. Introduction

Longitudinal studies are characterized by repeated measures over a period of time from each individual. Usually the subjects are assumed to be independent while the repeated measurements taken on each subject are correlated. The complication of longitudinal data analysis is partly due to the lack of a rich class of models such as the multivariate Gaussian for the joint distribution of the correlated responses (Liang and Zeger, 1986). Liang and Zeger (1986) introduced the generalized estimating equations (GEE) approach for analyzing longitudinal data in which a working correlation matrix for the responses of each individual is used. The GEE approach requires specification of only the first two moments of a subject's responses rather than the full specification of the joint distribution. The main advantage of the GEE estimation in longitudinal data analysis is that the estimators are consistent (asymptotically unbiased) even if the working correlation structure is misspecified. However, the GEE technique is asymptotic. If the sample size is small, the GEE method may produce biased estimators.

Under general conditions, maximum likelihood (ML) estimators are consistent. However, they are not unbiased generally. Cox and Snell (1968) provide general results for the first-order correction of bias of ML estimators of parameters under any distribution. Firth (1993) showed that the order $1/N$ bias of the ML estimator can be removed by introducing an appropriate bias term into the likelihood score function. The bias correction method of Cox and Snell (1968) is corrective and that of Firth (1993) is preventive. However, both methods are based on the likelihood score function. In this chapter we develop two analogous methods based on the generalized estimating function to obtain bias corrected estimates of the regression parameters in longitudinal data with specified working correlation matrix of the responses.

The GEE method estimates the regression parameters by constructing optimal linear combinations of the Pearson residuals. Thus this approach, like the weighted least squares method, does not possess any robust properties and is not sensitive to small deviations from the model assumptions. Wang, Lin and Zhu (2005) developed a robust version of the GEE approach and introduced a one-step bias correction technique to correct the asymptotic bias resulted from the asymmetric distribution of the Pearson residuals. By applying the robust estimation method of Wang, Lin and Zhu (2005), Qin, Zhu and Fung (2008) proposed two robustified estimating equations for the correlation parameters. The fist is the robustified version of the pseudolikelihood method (Huggins, 1993) and the second is the robustified version of the method by Wang and Carey (2004).

The method based on bias correction of the GEE estimates and that based on bias-reduced estimating equations are derived in Section 4.2. In section 4.3, the bias corrected and bias-reduced methods are applied to longitudinal binary and Poisson data. A simulation study is conducted in Section 4.4. Two examples are given in section 4.5 and a discussion follows in Section 4.6.

4.2. Estimates of the Regression Parameters Based on Bias-correction and Bias-reduction for Longitudinal Data

Let $y_n = (y_{n1}, \ldots, y_{nd})^T$ be the vector of responses with a $d \times p$ design matrix $X_n = (x_{n1}, \ldots, x_{nd})^T$ for the nth subject, $n = 1, \ldots, N$. Assume that the N subjects are independent while the repeated measurements y_{nj} taken on each subject are correlated. Define $\mu_n = \mathrm{E}(y_n|X_n) = (\mu_{n1}, \ldots, \mu_{nd})^T$ to be the expectation of y_n conditional on X_n and suppose $\mu_n = F(X_n\beta)$, where β is a $p \times 1$ vector of regression parameters of interest and F^{-1} is the link function. Assume that the variance of y_{nj} is given by $\phi v(\mu_{nj})$, where v is the variance function and ϕ is the overdispersion parameter. Note that, for binary data, F is the standard normal cumulative distribution for probit link and a standard logistic cumulative distribution for logit link. For Poisson data with log link, F is the exponential function.

The method of generalized estimating equations (GEE) proposed by Liang and Zeger (1986) for repeated measures does not specify a joint distribution of a subject's responses. Instead, it uses a common working correlation matrix for the longitudinal responses of each subject. Let $R(\rho)$ be a working correlation matrix completely specified by the parameter vector ρ of length q. Then $\phi W_n = \phi A_n^{1/2} R(\rho) A_n^{1/2}$ is the corresponding working covariance matrix, where $A_n(\beta) = \mathrm{diag}\{v(\mu_{nk})\}$, $k = 1, \ldots, d$, $n = 1, \ldots, N$. For given consistent estimates of ϕ and ρ, the GEE estimate of β, denoted by $\hat{\beta}$, is obtained by solving the generalized estimating equations

$$\sum_{n=1}^{N} D_n^T W_n^{-1}(y_n - \mu_n) = 0, \tag{4.2.1}$$

where $D_n = \frac{\partial \mu_n}{\partial \beta} = \Delta_n X_n$, $\Delta_n = \mathrm{diag}(f(x_{n1}^T\beta), \ldots, f(x_{nd}^T\beta))$ with $f = F'$, $n = 1, \cdots, N$.

Crowder (2001) modified the Gaussian estimation by forcibly decoupling $\mathrm{Var}(y_i)$ such that $\mathrm{Var}(y_i)$ depends only on ρ and ϕ, not on β. That is, the β in the mean $\mu_i(\beta)$ and the β in $\mathrm{Var}(y_i)$ are treated as distinct parameters. The estimating equations for the regression parameters by this modified Gaussian estimation approach are identical with the generalized estimating equations. Therefore, it is natural to regard the GEE function as a Gaussian score.

The left hand side of equation (4.2.1) which can be written as

$$U(\beta;\rho,\phi)=\sum_{n=1}^{N}(y_n-\mu_n)^T W_n^{-1}\frac{\partial\mu_n}{\partial\beta} \tag{4.2.2}$$

is the generalized estimating function for β given ρ and ϕ. Let $U(\beta;\rho,\phi)=(U_1,U_2,\ldots,U_p)$. For obtaining bias-corrected (Cox and Snell, 1968) and bias-reduced (Firth, 1993) GEE estimates, we treat U_i as if it were a likelihood score function for β_i, $i=1,\ldots,p$.

Now, define $\kappa_{ij}=\mathrm{E}(\partial U_i/\partial\beta_j)$ for $i,j=1,\ldots,p$. Further, define $\kappa_{ijl}=\mathrm{E}(\partial^2 U_i/\partial\beta_j\partial\beta_l)$, $\kappa_{ij,l}=\mathrm{E}\{(\partial U_i/\partial\beta_j)U_l\}$ and $\kappa_{ij}^{(l)}=\partial\kappa_{ij}/\partial\beta_l$ for $i,j,l=1,\ldots,p$. Then the Fisher information matrix analogue of order p for β is $I=\{-\kappa_{ij}\}$. Expressions for the quantities κ_{ij}, $\kappa_{ij}^{(l)}$ and κ_{ijl} are given in Appendix F.

Now, let $I^{-1}=\{-\kappa^{ij}\}$ be the inverse of I. Then, following Cox and Snell (1968) the first-order bias of $\hat{\beta}_s$ is given by

$$b_s(\beta)=\sum_i\sum_j\sum_l\frac{1}{2}\kappa^{si}\kappa^{jl}(\kappa_{ijl}+2\kappa_{ij,l}),\quad s=1,\ldots,p. \tag{4.2.3}$$

Following Cordeiro and Klein (1994) this can be expressed as

$$b_s(\beta)=\sum_{i=1}^{p}\kappa^{si}\sum_{j,l=1}^{p}\left[\kappa_{ij}^{(l)}-\frac{1}{2}\kappa_{ijl}\right]\kappa^{jl},\quad s=1,\ldots,p. \tag{4.2.4}$$

In matrix notation, the vector $b(\beta)=(b_1(\beta),\cdots,b_p(\beta))^T$ can be written as

$$b(\beta)=\mathrm{E}(\hat{\beta}-\beta)=I^{-1}A\mathrm{vec}(I^{-1}), \tag{4.2.5}$$

where $A=\{A^{(1)}|\cdots|A^{(p)}\}$ with $A^{(l)}=\{a_{ij}^{(l)}\}$ having its (i,j)th element defined by $a_{ij}^{(l)}=\kappa_{ij}^{(l)}-\frac{1}{2}\kappa_{ijl}$ for $l=1,\ldots,p$. Then, the bias corrected GEE estimate, using the method of Cox and Snell (1968), $\tilde{\beta}$, of β is given by

$$\tilde{\beta}=\hat{\beta}-b(\hat{\beta})=\hat{\beta}-\hat{I}^{-1}\hat{A}\mathrm{vec}(\hat{I}^{-1}), \tag{4.2.6}$$

where $\hat{I}^{-1}$ and $\hat{A}$ are the matrices I^{-1} and A computed at $\hat{\beta}$.

This approach is "corrective" rather than "preventive": the GEE estimate $\hat{\beta}$ is first calculated then corrected. Following the "preventive" method of Firth (1993), by introducing a bias term into the score function $U(\beta;\rho,\phi)$, the modified score function is

$$U^{*}(\beta;\rho,\phi) = U(\beta;\rho,\phi) - Ib(\beta).$$

The bias reduced GEE estimate, denoted by β^{*}, of β using the method of Firth (1993) is obtained by solving the modified score equation

$$U^{*}(\beta;\rho,\phi) = 0. \tag{4.2.7}$$

Following the GEE method, an iterative procedure for obtaining β^{*} can be described as in what follows.

Step 1: Choose an initial estimate $\beta^{(0)}$ of β (for example $\beta^{(0)}$ is obtained by the GEE method of estimation with independence working correlation matrix).

Step 2: Given β^{*} (at the first iteration $\beta^{*} = \beta^{(0)}$), the moment estimate of the overdispersion parameter is given by

$$\phi^{*} = \frac{1}{Nd}\sum_{n=1}^{N} Z_n^{*T} Z_n^{*T}, \text{ where } Z_n^{*T} = A_n^{-1/2}(\beta^{*})(y_n - \mu_n(\beta^{*})).$$

Step 3: Given β^{*} and ϕ^{*} obtained in Steps 1 and 2, calculate the moment estimates ρ^{*} of ρ of the working correlation matrix $R(\rho)$ (see Liang and Zeger, 1986 and Wang and Carey, 2003). For example, if the working correlation matrix is exchangeable, then the exchangeable correlation parameter ρ is estimated by

$$\rho^{*} = \frac{\sum_{n=1}^{N}\sum_{j\neq k} y_{nj}^{*} y_{nk}^{*}}{\phi^{*}(d-1)\sum_{n=1}^{N}\sum_{j=1}^{d} {y_{nj}^{*}}^{2}}, \text{ where } y_{nj}^{*} = (y_{nj} - \mu_{nj}(\beta^{*}))/\sqrt{v(\mu_{nj}(\beta^{*}))}.$$

See Wang and Carey (2003). If the working correlation matrix is AR(1), then the AR(1) correlation parameter ρ is estimated by

$$\rho^* = \frac{\sum_{n=1}^N \sum_{j=2}^d y_{nj}^* y_{n,j-1}^*}{\sum_{n=1}^N \{\sum_{j=2}^{d-1} {y_{nj}^*}^2 + ({y_{n1}^*}^2 + {y_{nd}^*}^2)/2\}}, \text{ where } y_{nj}^* = (y_{nj} - \mu_{nj}(\beta^*))/\sqrt{v(\mu_{nj}(\beta^*))}.$$

See Wang and Carey (2003).

Step 4: Given the working correlation matrix $R(\rho^*)$ obtained in Step 3, the estimate of β is updated according to the modified Fisher scoring formula

$$\beta_{j+1}^* = \beta_j^* + \left\{\sum_{n=1}^N D_n{}^T W_n{}^{-1} D_n\right\}^{-1}\Big|_{\beta=\beta_j^*} \left\{\sum_{n=1}^N D_n{}^T W_n{}^{-1}(Y_n - \mu_n(\beta)) - Ib(\beta)\right\}\Big|_{\beta=\beta_j^*},$$

where $D_n = \partial\mu_n/\partial\beta$ and $W_n = A_n(\beta)R(\rho^*)A_n(\beta)$.

Step 5: Iterate between steps 2 to 4 until a desired convergence criterion (for example, $\max|\beta_{j+1}^* - \beta_j^*| < 0.001$) is satisfied. At convergence, the estimate of β is denoted by β^* and the final estimates of ρ and ϕ are given by ρ^* and ϕ^* used in the last step of the iteration.

4.3. Application to binary and count data

4.3.1. Binary data.

For the vector of binary responses y_n, $n = 1, \ldots, N$, the variance function is given by $v(\mu) = \mu(1-\mu)$ and we consider the logit and probit link functions. For the probit link, $F = \Phi$ is the cumulative distribution function of the standard normal distribution. Thus, $f = F' = \phi$ is the density function of the standard normal distribution. Therefore, $\Delta_n = \text{diag}\{\phi(x_{n1}^T\beta), \ldots, \phi(x_{nd}^T\beta)\}$ and $\frac{\partial\Delta_n}{\partial\mu_n}$ is a $d^2 \times d$ dimensional sparse matrix with non-zero quantities $\phi'(\Phi^{-1}(\mu_{nj}))(\Phi^{-1})'(\mu_{nj})$ in the $[(j-1)d + j, j]$ term, $j = 1, \ldots, d$, $n = 1, \ldots, N$.

For the logit link, $F(x) = \frac{\exp(x)}{1+\exp(x)}$ is the standard logistic cumulative distribution function and $f(x) = F'(x) = \frac{\exp(x)}{(1+\exp(x))^2}$. Therefore, $\Delta_n = \text{diag}\{\frac{\exp(x_{n1}^T\beta)}{(1+\exp(x_{n1}^T\beta))^2}, \ldots, \frac{\exp(x_{nd}^T\beta)}{(1+\exp(x_{nd}^T\beta))^2}\}$

and $\frac{\partial \Delta_n}{\partial \mu_n}$ is a $d^2 \times d$ dimensional sparse matrix with non-zero quantities $(1 - 2\mu_{nj})$ in the $[(j-1)d + j, j]$ term, $j = 1, \ldots, d$, $n = 1, \ldots, N$.

Now, the matrix Δ_n is determined by the link function. Given a working correlation matrix $R(\rho)$ and the variance function $v(\mu) = \mu(1-\mu)$, the expressions on the right hand side of equations (F.1), (F.2) and (F.3) can be obtained. Thereafter, the bias corrected GEE (GEEBc) estimate $\tilde{\beta}$ is obtained from the formula (4.2.6) and the bias reduced GEE (BcGEE) estimate β^* is obtained by solving equation (4.2.7).

4.3.2. Count data.

For the vector of Poisson responses y_n, $n = 1, \ldots, N$, the variance function is given by $v(\mu) = \mu$ and if the log link function is used, then $f(x) = F'(x) = F(x) = \exp(x)$. Therefore, $\Delta_n = \mathrm{diag}\{\exp(x_{n1}^T\beta), \ldots, \exp(x_{nd}^T\beta)\}$ and $\frac{\partial \Delta_n}{\partial \mu_n}$ is a $d^2 \times d$ dimensional sparse matrix with 1 in the $[(j-1)d + j, j]$ term, $j = 1, \ldots, d$, $n = 1, \ldots, N$.

Now, given a working correlation matrix $R(\rho)$ and the variance function $v(\mu) = \mu$, the expressions on the right hand side of equations (F.1), (F.2) and (F.3) can be obtained. Thereafter, $\tilde{\beta}$ is obtained from the formula (4.2.6) and β^* is then obtained by solving equation (4.2.7).

4.4. Simulations

In this section we conduct a simulation study to compare the bias and efficiency properties of the estimates $\hat{\beta}$ (GEE estimates), $\tilde{\beta}$ (GEEBc estimates) and β^* (BcGEE estimates) of β for binary and Poisson data.

We first consider estimation of the regression parameters in a marginal model for correlated binary data with $p = 2$, $x_{ij}, j = 1, \ldots, d$, generated as uniform random variables in the interval $[-1, 1]$, $\beta = (0.5, 1.0)$ and a probit link function.

The correlated binary responses $y_i = (y_{i1}, \ldots, y_{id})^T$ are generated using the multivariate probit (MP) model (Ashford and Sowden, 1970 and Chaganty and Joe, 2004) in which the binary response is the indicator of the event that an unobserved latent variable exceeds a given threshold. Let $Z_i = (z_{i1}, \ldots, z_{id})^T$ be a d-dimensional vector of latent variables such that $Z_i = X_i\beta + \epsilon_i,\ i = 1, \ldots, N$, where

$$X_i = \begin{pmatrix} 1 & \cdots & 1 \\ x_{i1} & \cdots & x_{id} \end{pmatrix}^T.$$

The latent variable Z_i is assumed to follow a multivariate normal distribution with mean $X_i\beta$ and covariance $\Omega(\gamma)$, where γ is the latent correlation parameter. The binary response y_{ij} is given by

$$y_{ij} = \begin{cases} 1, & \text{if } z_{ij} > 0, j = 1, \ldots, d, \\ 0, & \text{otherwise,} \end{cases}$$

so that $P(y_{ij} = 1|X_i) = P(z_{ij} > 0) = \Phi(\beta_0 + \beta_1 x_{ij})$, where Φ is the standard normal distribution function.

Two sets of simulations are conducted: one with exchangeable latent correlation structure $\Omega(\gamma)$ with values of $\gamma = 0(0.1), \ldots, 0.9$ and the other with AR(1) laten correlation structure $\Omega(\gamma)$ with values of $\gamma = -0.9(0.1), \ldots, 0.9$.

The numbers of subjects taken are $N = 20, 30,$ 50 and 80 each subject having $d = 4$ observations. For each N, we simulate 5000 samples. We calculate bias($\hat{\beta}_0$) ($=\sum_{i=1}^{5000}(\hat{\beta}_{0i} - \beta_0)/5000$) and bias($\hat{\beta}_1$) ($=\sum_{i=1}^{5000}(\hat{\beta}_{1i} - \beta_1)/5000$) using each of the three methods GEE, GEEBc and BcGEE. Further we calculate relative efficiency (RE) of the estimates $\tilde{\theta}$ and θ^*, RE($\tilde{\theta}$) = MSE($\hat{\theta}$)/MSE($\tilde{\theta}$) and RE(θ^*) = MSE($\hat{\theta}$)/MSE(θ^*), where MSE($\hat{\theta}$) $= \sum_{i=1}^{5000}(\hat{\theta}_i - \theta)^2/5000$ and $\hat{\theta} = \hat{\beta}_0, \tilde{\beta}_0, \beta_0^*, \hat{\beta}_1, \tilde{\beta}_1$ and β_1^*.

Bias and efficiency properties of the estimates of β_0 and β_1 are very similar for all three methods. So we only give bias and efficiency results for the estimates of β_1. The bias results are summarized in Figure 4.1 for data generated using the exchangeable correlation structure and in Figure 4.2 for those using the AR(1) correlation structure.

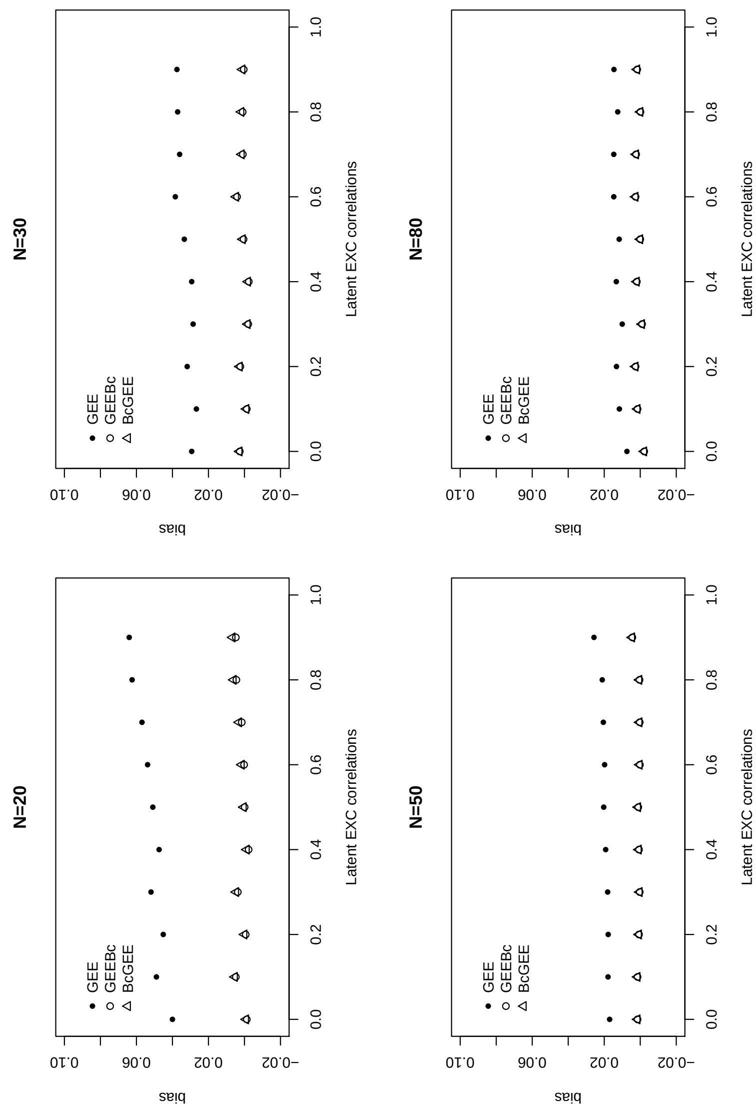

FIGURE 4.1. Biases of $\hat{\beta}_1$, $\tilde{\beta}_1$ and β_1^* with latent exchangeable correlations in MP model.

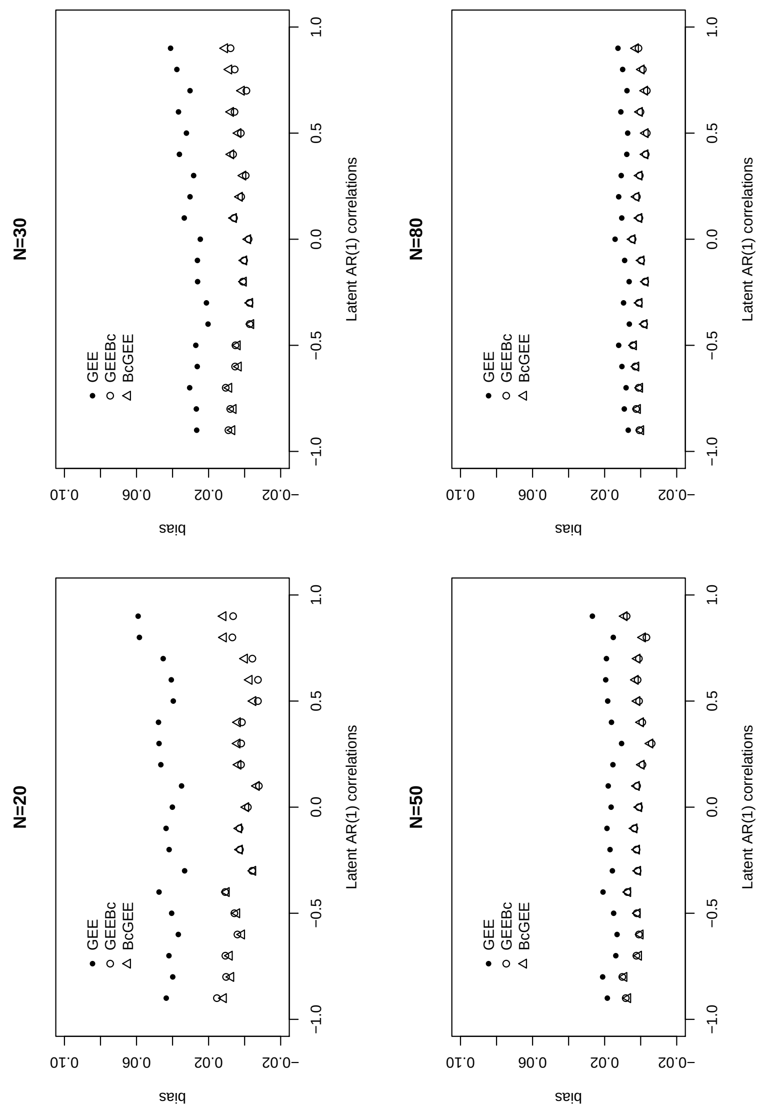

FIGURE 4.2. Biases of $\hat{\beta}_1$, $\tilde{\beta}_1$ and β_1^* with latent AR(1) correlations in MP model.

We first discuss the results in Figure 4.1. From the figure we see that $\hat{\beta}_1$ have largest biases for all values of the latent correlation considered. However, the biases of the estimates of $\tilde{\beta}_1$ and β_1^* seem to be similar. The difference between the biases of the estimates $\hat{\beta}_1$ and those of $\tilde{\beta}_1$ and β_1^* diminishes as the number of subjects increases. Very similar bias properties of all these estimates are observed from Figure 4.2 where data are generated using the AR(1) correlation structure.

We now compare efficiency of $\tilde{\beta}_1$ and β_1^* in relation to $\hat{\beta}_1$. These relative efficiency results are summarized in Figure 4.3 for data generated using the exchangeable correlation structure and in Figure 4.4 for data generated using the AR(1) correlation structure.

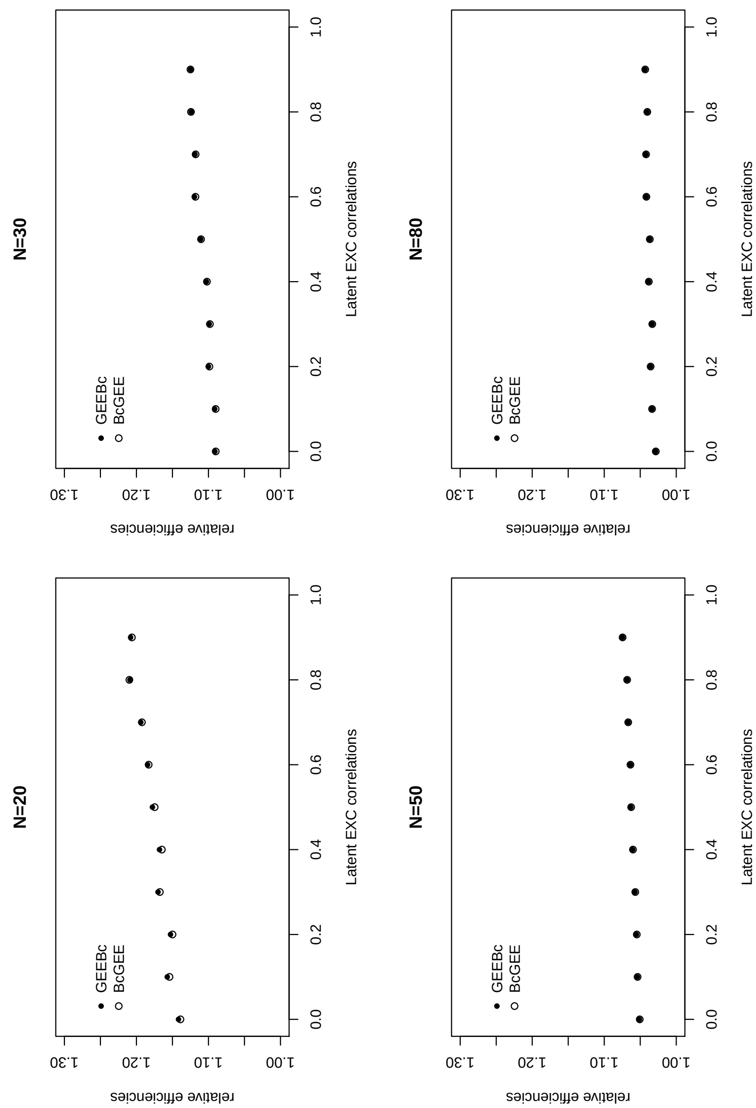

FIGURE 4.3. Relative efficiency of $\tilde{\beta}_1$ and β_1^* with latent exchangeable correlations in MP model.

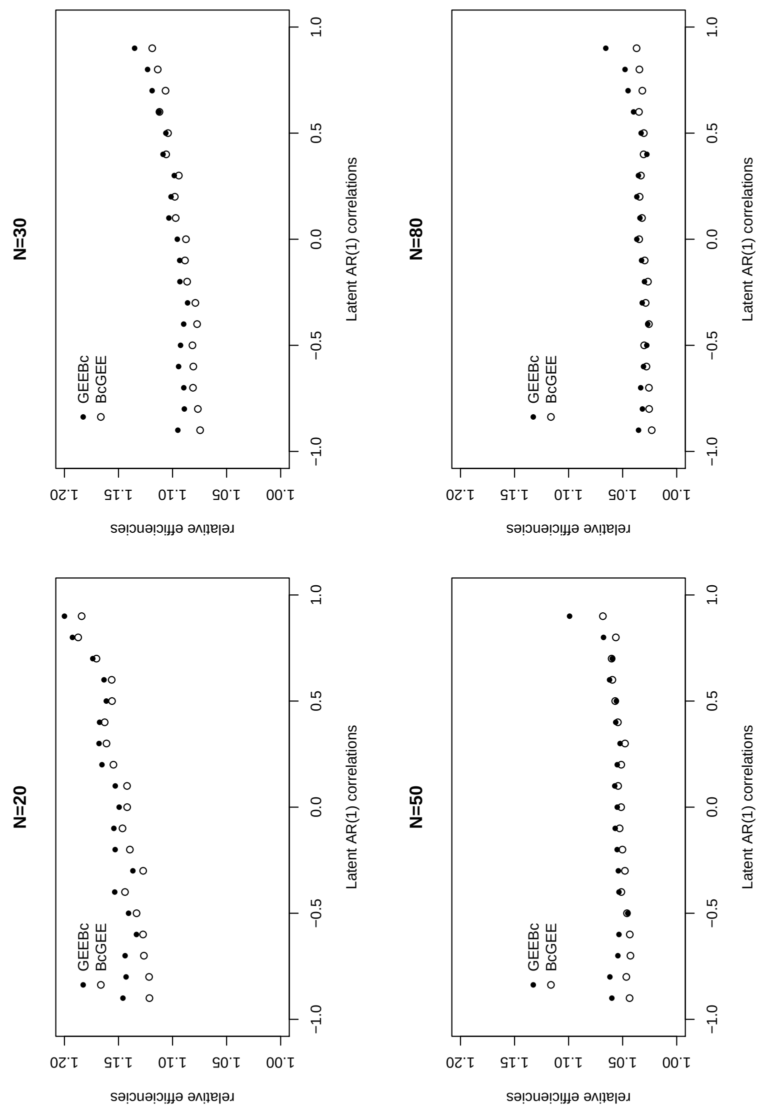

FIGURE 4.4. Relative efficiency of $\tilde{\beta}_1$ and β_1^* with latent AR(1) correlations in MP model.

From these figures it can be seen that the efficiencies of $\tilde{\beta}_1$ and β_1^* are very similar except in some cases when the data are generated using the AR(1) correlation structure in which $\tilde{\beta}_1$ is slightly more efficient. In general both estimates are more efficient than the GEE estimates $\hat{\beta}_1$ for small number of subjects ($N = 20$, $N = 30$). As the number of subjects increases ($N = 80$), relative efficiencies of $\tilde{\beta}_1$ and β_1^* become closer to 1 indicating the benefit of the bias correction procedure for small clusters situation.

We now study the performance of $\hat{\beta}_1$, $\tilde{\beta}_1$ and β_1^* for longitudinal Poisson data. For the generation of longitudinal Poisson data, we consider N subjects, each with $d = 4$ repeated responses such that $\mu_{ij} = \exp(\beta_0 + \beta_1 x_{ij})$ and $\text{Var}(y_{ij}) = \mu_{ij}$, where $\beta_0 = 0$, $\beta_1 = 0.5$, x_{ij} is generated from a uniform distribution on $[j-1, j]$, $j = 1, \ldots, 4$, $i = 1, \ldots, N$. EXC and AR(1) Poisson data y_{ij} are generated using the method of Yahav and Shmueli (2011).

For large N, bias and efficiency properties of $\hat{\beta}_1$, $\tilde{\beta}_1$ and β_1^* are very similar. Also, for data generated from AR(1) correlation structure these properties of the three estimates do not differ much even for small N. In general biases become closer to 0 and efficiencies become closer to 1 for all three estimates. These results are not presented here. For small N and when data are generated from the exchangeable correlation structure, the GEE estimates $\hat{\beta}_1$ show some bias, whereas the bias of the other two become closer to 0 (see Figure 4.5). In this case efficiencies of $\tilde{\beta}_1$ and β_1^* are almost identical and increase as the true exchangeable correlation increases.

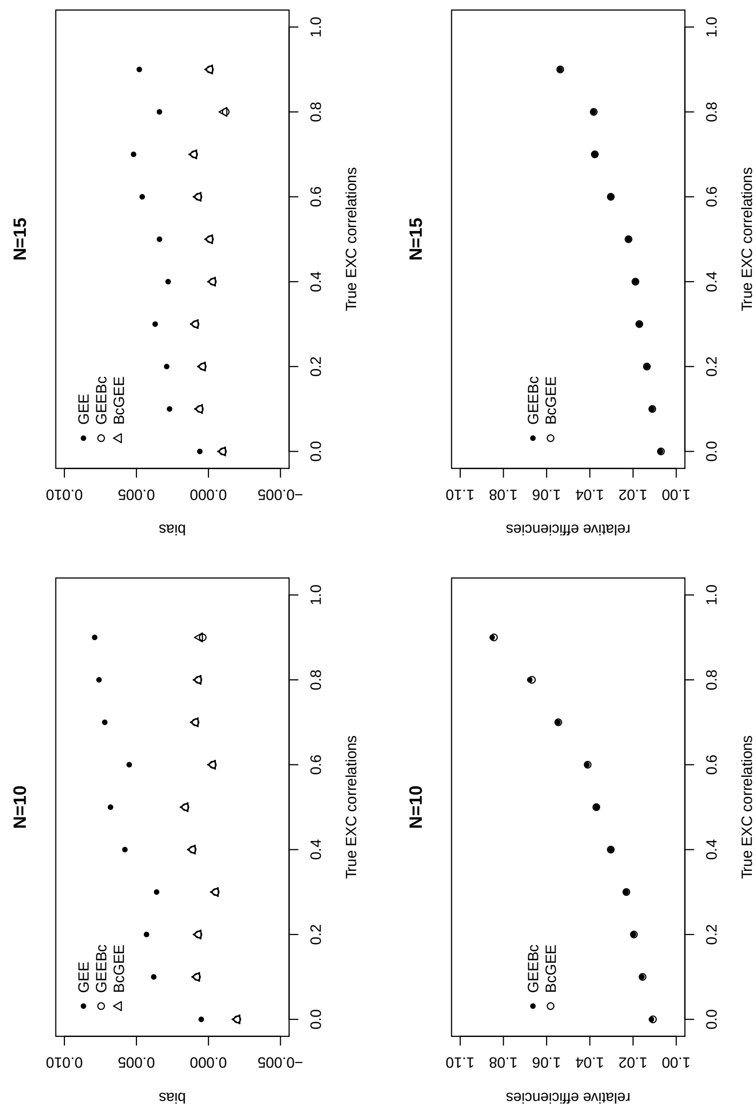

FIGURE 4.5. Biases and relative efficiencies of $\hat{\beta}_1$, $\tilde{\beta}_1$ and β_1^* for exchangeable Poisson data.

4.5. Examples

Example 1:

As a first example we consider a clinical trial on cerebrovascular deficiency with crossover design. The data set is from Diggle, Liang and Zeger (1994). The purpose of this crossover trial is to compare an active drug (A) and a placebo (B). A total of 67 patients were enrolled into the clinical trial of which 34 patients received the active drug (A) followed by placebo (B) and another 33 patients were treated in the reverse order. The response variable is defined to be 0 for an abnormal and 1 for a normal electrocardiogram reading. Conceptually, the 2×2 crossover trial can be viewed as a longitudinal study with 2 observations for each patient. The two major covariates, period (x_{i1}) and treatment (x_{i2}), are both time-dependent. They are coded as

$$x_{i1} = \begin{cases} 1, & \text{period 2} \\ 0, & \text{period 1,} \end{cases}$$

and

$$x_{i2} = \begin{cases} 1, & \text{active drug (A)} \\ 0, & \text{placebo (B)} \end{cases}$$

respectively. An analysis of full logistic regression model by Diggle et al. (1994) shows little support for a treatment-by-period interaction. Therefore, we consider the logistic regression model

$$\text{logit } \Pr(Y_{ij} = 1) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}.$$

The GEE estimates of the regression parameters β_0, β_1 and β_2 are 0.6659(0.2879), -0.2950(0.2311) and 0.5689(0.2327) respectively. The GEEBc estimates are 0.6527(0.2879), -0.2883(0.2312) and 0.5557(0.2328) respectively. The preventive BcGEE estimation gives bias-reduced estimates 0.6527(0.2865), -0.2876(0.2296) and 0.5556(0.2310) respectively. The standard errors of the estimates are given using the sandwich formula

TABLE 4.1. A subset of the 2×2 crossover trial data from Diggle et al. (1994).

	Responses					Period	
Group	(1, 1)	(0, 1)	(1, 0)	(0, 0)	Total	1	2
AB	5	0	3	2	10	8	5
BA	4	2	2	2	10	6	6

in the GEE estimation. As in the simulation study, the GEEBc and BcGEE estimates are almost identical. There seems to be some difference between the GEE estimates and the two bias corrected GEE estimates, although the difference is minimal, as the sample size of 67 is quite large.

Example 2:

To check what happens in small sample size situations we investigated many samples of the data in Example 1 of size 20 (10: 10) of which one sample is given in Table 4.1. For this sample, the GEE estimates of the regression parameters β_0, β_1 and β_2 are 0.5381(0.5777), -0.6694(0.5465) and 0.6694(0.5465) respectively. The GEEBc estimates are 0.4974(0.5777), -0.6181(0.5469) and 0.6181(0.5469) respectively and the BcGEE estimates are 0.5003(0.5705), -0.6208(0.5389) and 0.6208(0.5389) respectively. As can be seen, again, there is not much difference between the GEEBc and BcGEE estimates. However, significant differences between the GEE estimates and those by the other two methods is observed. This property was also observed for all the samples investigated.

4.6. Discussion

In this chapter we obtain two bias corrected GEE estimates of the regression parameters in longitudinal data. One of these, GEEBc, is based on correcting the bias of the GEE estimates following Cox and Snell (1968). The other, BcGEE, is based on correcting the GEE following a method by Firth (1993). The performance in terms of bias and efficiency of both of these estimates are very similar and both show superior performance in terms of bias and efficiency compared to the GEE estimates for small samples. An example provided confirms these findings.

A SAS macro can be easily written to obtain the GEEBc and BcGEE estimates of the regression parameters in longitudinal data. We have written a SAS macro (%BcGEE) which can be implemented to analyze binary and Poisson longitudinal data.

CHAPTER 5

Effects of Variance Function on Estimation Efficiency

5.1. Introduction

The main advantage of the GEE method of estimation in longitudinal data analysis is that the estimators are consistent even if the working correlation structure is misspecified. Although misspecification of the correlation structure does not affect consistency of the estimates of the regression parameters, it does reduce the efficiency of the regression parameter estimates (Wang and Carey, 2003). As discussed by Wang and Zhao (2007) the GEE approach pays attention to correctly modelling the working correlation matrix $R(\rho)$, however, it treats the variance function to be of a known form obtained from the generalized linear models (GLM). In practice the distribution of the data may not be from a GLM and therefore the chosen variance function may be wrong. Wang and Lin (2005) investigated the impacts of misspecifying the variance function on estimators of the regression parameters. They showed that if the variance function is misspecified, the correct choice of the correlation structure may not necessarily improve estimation efficiency. The purpose of this chapter is to study the problem of estimating the parameters of the variance function assuming that the form of the variance function is known and then the effect of a misspecified variance function on the estimates of the regression parameters.

In the framework of a GLM the variance function for normal data is $v(\phi, \mu) = \phi$. For over-dispersed count data the variance function is $v(\phi, \mu) = \phi\mu$, where ϕ is the over-dispersed parameter. A popular variance function for count data is that given by the negative binomial, namely $V(c, \mu) = \mu(1 + c\mu) = \mu + c\mu^2$ (see, for example, Paul and Islam, 1995 and Paul and Banerjee, 1998). More general variance functions for count data include $v(\gamma, \mu) = c\mu^{\gamma}$, where $1 \leq \gamma \leq 2$, and $v(\gamma, \mu) = \gamma_1\mu + \gamma_2\mu^2$ (Paul and Plackett, 1978), where the parameters (c, γ) and $(\gamma_1.\gamma_2)$ need to be estimated. All the other variance functions, namely, $v(\phi, \mu) = \phi$, $v(\phi, \mu) = \phi\mu$ and $V(c, \mu) = \mu + c\mu^2$ are special cases of these two variance functions. The variance function $v(\gamma, \mu) = c\mu^{\gamma}$ can also be used for continuous data (See Davidian and Giltiman, 1995).

In this chapter we propose a GEE approach to estimate the parameters of the chosen variance function. The estimation method borrows the idea of Davidian and Carroll (1987) by solving a non-linear regression problem where residuals are regarded as the responses and the variance function is regarded as the regression function. We investigate the impact of misspecification of variance functions on efficiency of the estimates of the regression parameters and compare our method with the modified pseudo-likelihood approach of Wang and Zhao (2007) by simulations.

In Section 2 we review the modified pseudo-likelihood approach (Wang and Zhao, 2007) to estimate the parameters of the chosen variance function. In Section 3 we develop the GEE approach. A simulation study is conducted in Section 4.

5.2. Modified pseudo-likelihood approach (Wang and Zhao, 2007)

Wang and Zhao (2007) used a modified pseudo-likelihood approach in which the parameters of the variance function are estimated using the Gaussian likelihood (Whittle, 1961) and the regression parameters β are estimated using the GEE approach.

The Gaussian estimation method (Whittle, 1961) uses the normal log likelihood without assuming that the data are normally distributed. For the longitudinal data

$Y_1, \ldots, Y_N$ the working log likelihood is

$$G(\theta) = G(\beta, \rho, \tau) = -\frac{1}{2}\sum_{i=1}^{N}[\log\{\det(2\pi V_i)\} + (Y_i - \mu_i)^{\mathrm{T}} V_i^{-1}(Y_i - \mu_i)],$$

where $V_i = A_i R(\rho) A_i$ and $A_i = \mathrm{diag}\{\sqrt{v(\tau, \mu_i)}\}$, where τ is the vector of parameters involved in the variance function. For the estimation of τ Wang and Zhao (2007) sets $R(\rho) = I_d$ in $G(\theta)$, where I_d is a d dimensional identity matrix. Given an estimate $\hat{\beta}$ of β, the estimate of τ is obtained by maximizing $G_0(\tau, \hat{\beta})$ with respect to τ, where

$$G_0(\tau, \hat{\beta}) = -\frac{1}{2}\sum_{i=1}^{N}[\log\{\det(\hat{A}_i^2)\} + (Y_i - \hat{\mu}_i)^{\mathrm{T}} \hat{A}_i^{-2}(Y_i - \hat{\mu}_i)], \tag{5.2.1}$$

where $\hat{\mu}_i$ and $\hat{A}_i$ are evaluated at $\beta = \hat{\beta}$.

Note that if we consider the variance function $v(\gamma, \mu) = c\mu^{\gamma}$ then from (5.2.1), given $\hat{\beta}$ and γ, the estimate $\hat{c}$ of the scale parameter c (> 0) is given by

$$\hat{c} = \frac{1}{Nd}\sum_{i=1}^{N}\hat{Z}_i^{\mathrm{T}}\hat{Z}_i, \tag{5.2.2}$$

where $\hat{Z}_i = \mathrm{diag}(\mu_i^{-\gamma/2}(\hat{\beta}))$. Now, given $\hat{\beta}$ and $\hat{c}$ (which involves γ and $\hat{\beta}$), $\hat{\gamma}$ is obtained by maximizing G_0 with respect to γ. In G_0, the specific choice of the independence correlation $R(\rho) = I_d$ guarantees that the Gaussian estimating function for γ is unbiased (see Wang and Zhao, 2007). That is, the scale parameter c is playing the same role as σ^2 of ordinary least square (OLS) theory and its estimate is given by the mean residual sum of squares.

Wang and Zhao (2007) claimed that the estimate $\hat{\gamma}$ by maximizing $G_0(\gamma, c, \hat{\beta})$ is consistent. However, the consistency property is based on the assumption that $\hat{A}_i = A_i$ which depends on the correct estimate of γ.

If we combine the modified pseudo-likelihood approach for estimating the variance parameter γ and the GEE estimation of the regression parameter β, this procedure can be described as what follows.

Step 1: The initial estimate β^0 of β is obtained by the GEE method of estimation with independence working correlation assuming a usual variance function that does not involve γ. This GEE estimation ensures that β^0 is a consistent estimate of β.

Step 2: Given estimate $\hat{\beta}$ (at the first iteration $\hat{\beta} = \beta^0$), the estimate $\hat{\gamma}$ and $\hat{c}$ are obtained by maximizing $G_0(\gamma, c, \hat{\beta})$ given in (5.2.1) with respect to $\tau = (\gamma, c)$.

Step 3: Given $(\hat{\beta}, \hat{\tau})$ obtained in steps 1 and 2, the correlation parameter ρ of the chosen working correlation matrix is estimated by the method of moments (Liang and Zeger, 1986 and Wang and Carey, 2003).

Step 4: Given $(\hat{\tau}, \hat{\rho})$ obtained in steps 2 and 3, the estimator of the regression parameter $\hat{\beta}$ is updated according to the modified Fisher scoring formula for β in GEE method of estimation

$$\hat{\beta}^{(j+1)} = \hat{\beta}^{(j)} + \left\{\sum_{i=1}^{N} \hat{D}_i^T \hat{W}_i^{-1} \hat{D}_i\right\}^{-1} \left\{\sum_{i=1}^{N} \hat{D}_i^T \hat{W}_i^{-1} (Y_i - \hat{\mu}_i)\right\}, j = 1, 2, \cdots,$$

where $\hat{D}_i = \partial\mu_i / \partial\beta\big|_{\hat{\beta}^{(j)}}$, $\hat{\mu}_i = \mu_i(\hat{\beta}^{(j)})$ and $\hat{W}_i = A_i(\hat{\gamma}, \hat{\beta}^{(j)}) R(\hat{\rho}) A_i(\hat{\gamma}, \hat{\beta}^{(j)})$.

Step 5: Iterate between steps 2 to 4 until a desired convergence criterion (for example $\max|\hat{\beta}^{(j+1)} - \hat{\beta}^{(j)}| < 0.001$) for β is satisfied. At convergence, the final estimates of τ and ρ are those used in step 2 and 3 of the iteration.

5.3. Estimating parameters of the variance function using generalized estimating equations

Davidian and Carroll (1987) developed a general theory for variance function estimation in heteroscedastic regression models. Heteroscedasticity in regression analysis is modelled as a function of the covariates and other parameters through a variance function. Some of the variance functions discussed by Davidian and Carroll (1987) include $v(c, \gamma, \mu) = c\mu^\gamma$ and $v(\gamma_1, \gamma_2, \mu) = \gamma_1\mu + \gamma_2\mu^2$.

In the usual nonlinear regression setup let $\hat{\beta}_*$ be a preliminary estimator for β and denote the residuals by $r_i = y_i - \mu_i(\hat{\beta}_*)$. Davidian and Carroll (1987) suggested estimating the parameters of the variance function by solving a nonlinear regression problem in which the "responses" are r_i^2 and the "regression function" is $v(x_i, \hat{\beta}_*, \tau)$. The motivation of this method is that the squared residuals have approximate expectation $v(x_i, \beta, \tau)$. For normal data, the squared residuals have approximate variance $v^2(x_i, \beta, \tau)$ and thus the generalized least square method can be applied to estimate τ (see Davidian and Carroll, 1987).

For longitudinal data analysis, we consider a variance function $v(\tau, \mu)$, where $\tau = (c, \gamma)$ in $v(c, \gamma, \mu) = c\mu^\gamma$ and $\tau = (\gamma_1, \gamma_2)$ in $v(\gamma_1, \gamma_2, \mu) = \gamma_1\mu + \gamma_2\mu^2$. Other variance functions can be similarly treated. let $\hat{\beta}$ be the estimate of β obtained using the GEE method. Then, the parameter τ can be consistently estimated by solving the estimating equation

$$\sum_{i=1}^{N} \Big(\partial v(\tau, \mu_i(\hat{\beta}))/\partial\tau\Big)^T V_i^{-1}(\tau)(r_i^2 - v(\tau, \mu_i(\hat{\beta}))) = 0$$

for τ, where $V_i(\tau) = B_i Cr B_i$, $B_i = \sqrt{\mathrm{Var}(r_i^2)}$, $Cr = \mathrm{Corr}(r_i^2)$ is the correlation matrix for the "response" vector r_i^2. The structure of Cr might be different from that of the correlation matrix $R(\rho)$. Therefore, we are not trying to model the structure of Cr. Instead, we use the identity matrix for Cr since it does not affect the consistency of the estimate of τ. Further, we use $\mathrm{diag}\{v^2(\tau, \mu_i(\hat{\beta}))\}$ to approximate the variance of r_i^2. Note that this is the true value of $\mathrm{Var}(r_i^2)$ if the data are normally distributed. Therefore, the above estimating equation for τ becomes

$$\sum_{i=1}^{N} E_i^T V_i^{-1}(\tau)(r_i^2 - v(\tau, \mu_i(\hat{\beta}))) = 0, \tag{5.3.1}$$

where $E_i = \partial v(\tau, \mu_i(\hat{\beta}))/\partial\tau$, $V_i(\tau) = \mathrm{diag}\{v^2(\tau, \mu_i(\hat{\beta}))\}$.

Thus, given estimates $\hat{\beta}$ the GEE estimate of τ is obtained by the modified (Fisher scoring) iterative procedure:

$$\hat{\tau}^{(j+1)} = \hat{\tau}^{(j)} + \left\{\sum_{i=1}^{N} \tilde{E}_i^T \tilde{V}_i^{-1} \tilde{E}_i\right\}^{-1} \left\{\sum_{i=1}^{N} \tilde{E}_i^T \tilde{V}_i^{-1} (r_i^2 - v(\hat{\tau}^{(j)}, \mu_i(\hat{\beta}))\right\}, j = 1, 2, \cdots, \quad (5.3.2)$$

where $\tilde{E}_i = \partial v(\tau, \mu_i(\hat{\beta}))/\partial\tau\Big|_{\tau=\hat{\tau}^{(j)}}$ and $\tilde{V}_i = \text{diag}\{v^2(\hat{\tau}^{(j)}, \mu_i(\hat{\beta}))\}$.

Now, an iterative procedure for obtaining the estimates τ, ρ and β can be described as follows:

Step 1: Choose initial estimates β^0 for β (for example, β^0 are obtained by the GEE method of estimation with independence working correlation assuming a usual variance function that does not involve τ).

Step 2: Given $\beta = \tilde{\beta}$ (at the first iteration $\tilde{\beta} = \beta^0$, obtain an estimate $\tilde{\tau}$ of τ using the iteration procedure (5.3.2).

Step 3: Given $\beta = \tilde{\beta}$, $\tau = \tilde{\tau}$ obtained in steps 1 and 2, the estimate $\tilde{\rho}$ of the correlation parameter ρ is obtained by the method of moments (Liang and Zeger, 1986 and Wang and Carey, 2003).

Step 4: Calculate $\tilde{A}_i = A_i(\tilde{\tau}, \tilde{\beta})$, $\tilde{\mu}_i = \mu_i(\tilde{\beta})$ and $\tilde{D}_i = D_i(\tilde{\beta})$, $1 \leq i \leq N$. Then update the estimate of β as

$$\hat{\beta} = \tilde{\beta} + \left\{\sum_{i=1}^{N} \tilde{D}_i^T \tilde{\Sigma}_i^{-1} \tilde{D}_i\right\}^{-1} \left\{\sum_{i=1}^{N} \tilde{D}_i^T \tilde{\Sigma}_i^{-1} (Y_i - \tilde{\mu}_i)\right\}.$$

Step 5: Iterate between steps 2 to 4 until a desired convergence criterion (for example $\max|\hat{\beta} - \tilde{\beta}| < 0.001$) for β is satisfied. At convergence, the final estimates of τ and ρ are given by $\hat{\tau} = \tilde{\tau}$ and $\hat{\rho} = \tilde{\rho}$ used in the last step of iteration.

5.4. Simulations

In this section a simulation study is conducted. The purpose of the simulation is a) to compare efficiency of the estimates of the regression parameters β using the

GEE method (M_1), the pseudo-likelihood method of Wang and Zhao (2007) (M_2) and the new method proposed in this chapter (M_3); b) to investigate the effects of misspecification of variance functions on the estimating efficiency of the regression parameters β; and c) to investigate the effects of misspecified functional form of the variance function on the efficiency of the estimates of the regression parameters β. Simulations are conducted for data generated from normal populations.

We use a simulation design similar to Wang and Zhao (2007). We generate multivariate normal data using the linear model, $\mu_{ij} = \beta_0 + \beta_1 x_{ij}$ with $\beta_0 = 0.0$, $\beta_1 = 1.0$ and x_{ij} generated from a uniform distribution on $(j, j+1)$. Further, we use the variance function, $\mathrm{Var}(y_{ij}) = \phi\mu_{ij}^{\gamma}$ with $\phi = 1$ and values of $\gamma = 1.5$.

We generated data for $N = 100$ subjects, each with $d = 4$ repeated measures. In order to study the effect of misspecification of the correlation structure we generated repeated measures data with AR(1) and EXC (exchangeable) correlation structures with the correlation parameter ρ taking values of -0.9 to 0.9 by an increment of 0.1. Estimation of the regression parameters are also obtained using the two correlation structures and with $\gamma = 0, 1.5, \hat{\gamma}$, 2.5, 3.5.

Results of efficiency of $\hat{\beta}_1$ are given in Figure 5.1. Results in the graph reveal some interesting conclusions, namely, (i) Efficiency of $\hat{\beta}_1$ by the methods M_2 and M_3 is better than that by the method M_1 (the GEE method). Performances of the methods M_2 and M_3 are similar. Further, it appears that efficiency of $\hat{\beta}_1$ using the true value of γ (=1.5, in this case) is very similar to that by the method M_2 or M_3. This assures the reliability of both the methods M_2 and M_3. However, departure of the value of γ from the true value results in loss of efficiency. The loss of efficiency becomes larger and larger as the departure of the value of γ from the true value increases.

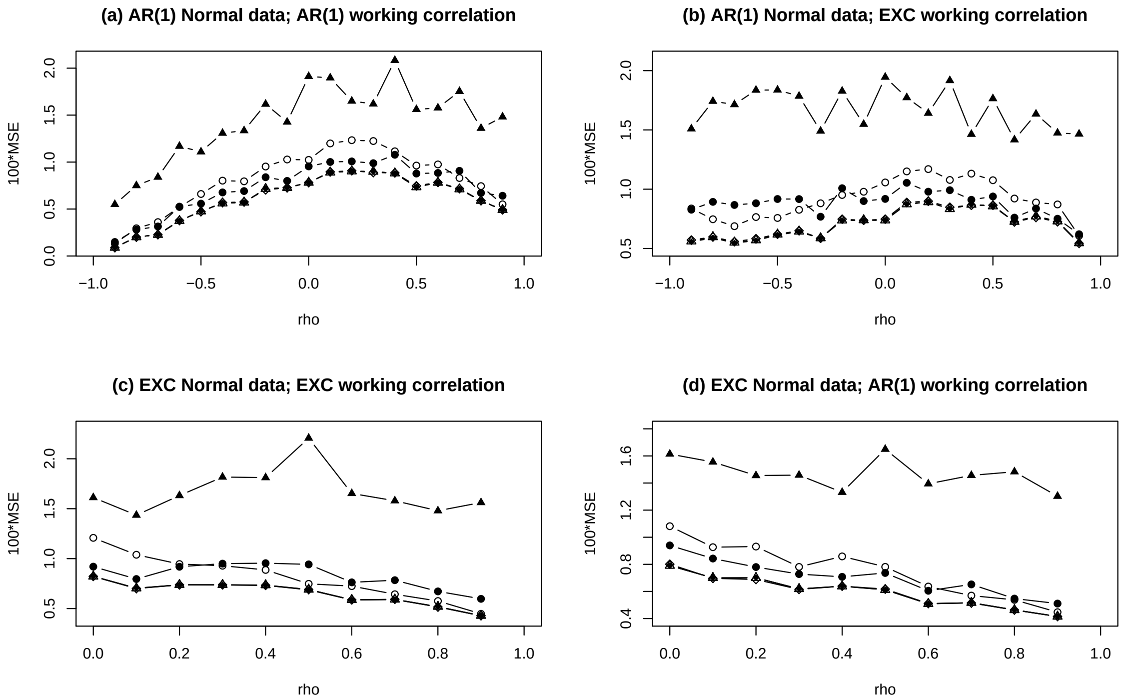

FIGURE 5.1. Comparison of MSE of $\hat{\beta}_1$ for longitudinal normal data by fixing γ in the power function at 0 ($\circ$), 1.5($\diamond$), 2.5($\bullet$), 3.5($\blacktriangle$) or estimating γ by the proposed method ($\triangle$) and the pseudolikelihood method ($\blacklozenge$). Data are generated from either AR(1) or EXC correlation structures. The working correlation structure is either AR(1) or EXC. The true values are $\beta_0 = 0, \beta_1 = 1$ and $\gamma = 1.5$.

To investigate the effects of misspecified functional form of the variance function on the efficiency of the estimates of the regression parameters β we extended the simulation study. We first generate multivariate normal data using the variance function $\text{Var}(y_{ij}) = \phi\mu_{ij}^{\gamma}$ (the power function). We then estimate β using the two methods M_2 and M_3. For each method two sets of estimates are obtained, one by the power function as the working variance function and the other by using $v(\gamma_1, \gamma_2, \mu) = \gamma_1\mu + \gamma_2\mu^2$ (the Bartlett function). The results for MSE are given in Figure 5.2. This simulation was repeated by generating data using the Bartlett variance function. The results for MSE are given in Figure 5.3.

For normal data there does not appear to be any difference in efficiency of $\hat{\beta}$ between the two methods in any scenario (Figure 5.2 (a): data simulated using the power variance function and AR(1) correlation structure, but estimated using working power variance function, working Bartlett variance function and working AR(1) correlation structure; Figure 5.2 (b): data simulated using the power variance function and AR(1) correlation structure, but estimated using working power variance function, working Bartlett variance function and working exchangeable correlation structure; Figure 5.2 (c): data simulated using the power variance function and exchangeable correlation structure, but estimated using working power variance function, working Bartlett variance function and working exchangeable correlation structure; Figure 5.2 (d): data simulated using the power variance function and exchangeable correlation structure, but estimated using working power variance function, working Bartlett variance function and working AR(1) correlation structure; Figure 5.3 (a): data simulated using the Bartlett variance function and AR(1) correlation structure, but estimated using working power variance function, working Bartlett variance function and working AR(1) correlation structure; Figure 5.3 (b): data simulated using the Bartlett variance function and AR(1) correlation structure, but estimated using working power variance function, working Bartlett variance function and working exchangeable correlation structure; Figure 5.3 (c): data simulated using the Bartlett variance function and exchangeable correlation structure, but estimated using working power variance function, working Bartlett variance function and working exchangeable correlation structure; Figure 5.3 (d): data simulated using the Bartlett variance function and exchangeable correlation structure, but estimated using working power variance function, working Bartlett variance function and working AR(1) correlation structure).

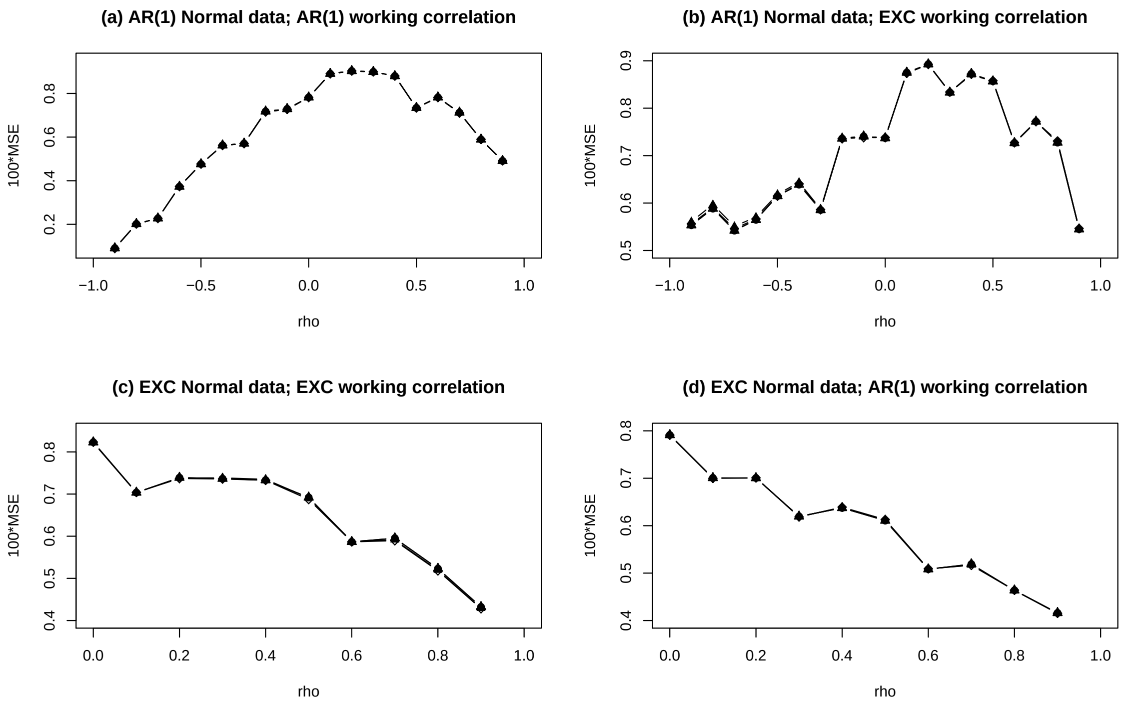

FIGURE 5.2. Comparison of MSE of $\hat{\beta}_1$ for longitudinal normal data generated using the power variance function $\gamma_1 \mu^{\gamma_2}$. In estimation the power variance function is used where the parameters are estimated by the pseudolikelihood method ($\blacktriangle$) and the proposed method ($\diamond$) or the Bartlett function is used where the parameters are estimated by the pseudolikelihood method ($\bullet$) and the proposed method ($\triangle$). Data are generated from either AR(1) or EXC correlation structures. The working correlation structure is either AR(1) or EXC. The true values are $\beta_0 = 0, \beta_1 = 1$ and the parameters in the power function $\gamma_1 = 1$ and $\gamma_2 = 1.5$.

Comparative properties of these methods for data from other distributions (for example, from over-dispersed Poisson, over-dispersed binomial) will be investigated in a future study and the overall results will be submitted to a journal in statistics.

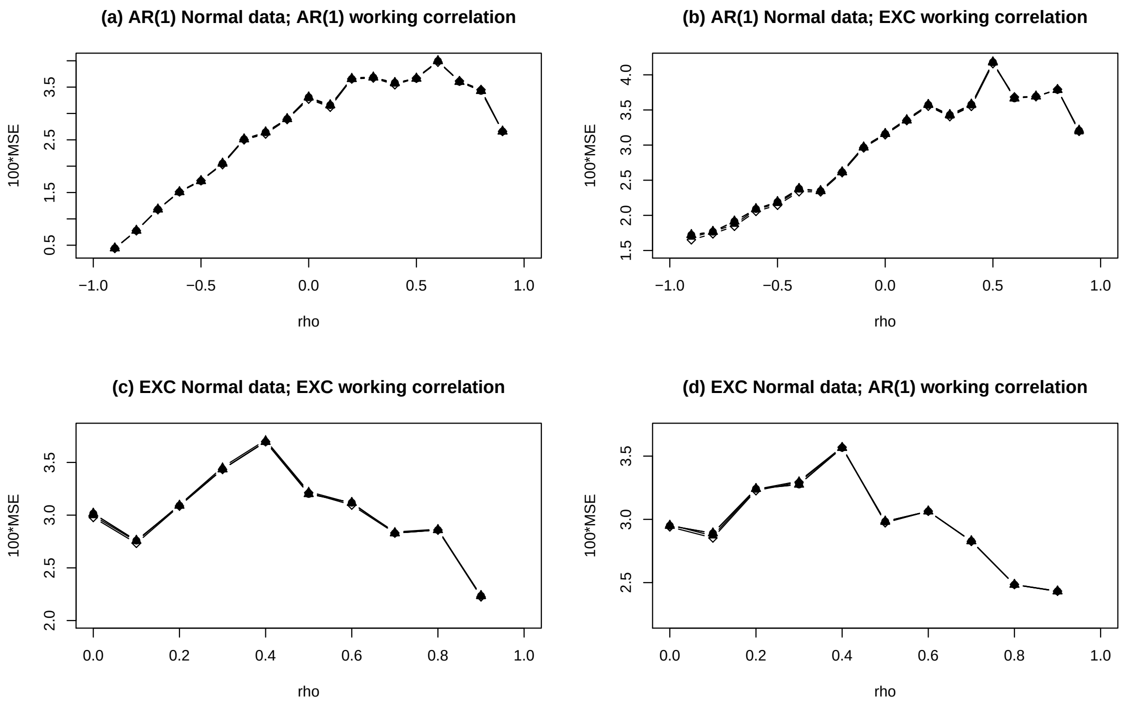

FIGURE 5.3. Comparison of MSE of $\hat{\beta}_1$ for longitudinal normal data generated using the Bartlett variance function $\gamma_1\mu^{\gamma_2}$. In estimation the power variance function is used where the parameters are estimated by the pseudolikelihood method (▲) and the proposed method ($\diamond$) or the Bartlett function is used where the parameters are estimated by the pseudolikelihood method ($\bullet$) and the proposed method ($\triangle$). Data are generated from either AR(1) or EXC correlation structures. The working correlation structure is either AR(1) or EXC. The true values are $\beta_0 = 0, \beta_1 = 1$ and the parameters in the power function $\gamma_1 = 1$ and $\gamma_2 = 2.5$.

CHAPTER 6

Conclusions and Future Research

In this last chapter, we first summarize the Gaussian estimation approach for longitudinal binary data, bias correction approach in GEE estimation and GEE estimation of variance parameters to improve estimation efficiency. Then we conclude with related future research.

We revisit the traditional Gaussian estimation approach to analyze longitudinal binary data by using a working correlation structure. For example, the choice of this robust working correlation, general autocorrelation, can ensure the asymptotic unbiasedness of the Gaussian estimating equations no matter the true correlation structure is exchangeable or AR(1). This distribution-free method has the advantage of higher efficiency of the β-estimator. But one limitation of this method is that the estimating equations are unbiased asymptotically. For example, if the working general autocorrelation structure is chosen and the true correlation structure is not exchangeable, AR(1) or general autocorrelation, then the estimating equations may not be asymptotically unbiased. Simulations show that the choice of the working correlations does not affect the estimation consistency and efficiency a lot. This may suggest that the asymptotic unbiasedness of the Gaussian estimating equations or consistency of the β estimator does not depend on the choice of the working correlations. Theoretical justifications of this result are left as a future research.

When the sample size (number of subjects and observation times) is small, we use bias correction techniques for ML estimation by treating the GEE function as a likelihood score to reduce the bias of the β estimator. The techniques include a corrective approach in which the GEE estimator is first calculated then corrected and a preventive approach by introducing a bias term into the GEE function. Simulations show that both approaches can reduce the GEE bias substantially when the sample size is small and the performance of these two methods are similar in terms of bias and efficiency. However, it is not very often that the sample size is small in longitudinal data analysis. Therefore, one more important future area of research is to investigate the bias correction problem when the covariates are time dependent. This problem will be described in detail in section 6.1.

The misspecification of the variance function has more effects on the efficiency of the GEE estimators than the misspecification of working correlation structures. Therefore, estimation of the variance function is more important than the choice of the working correlation structure in the GEE method. We use the GEE method to estimate the variance parameters by solving a non-linear regression problem in which the residuals are regarded as responses and the variance function is regarded as the regression function. However, the form of the variance function is unknown in practice. More research is needed to investigate the variance function selection criteria.

6.1. Marginal regression analysis of longitudinal data with time-dependent covariates

The main advantage of the GEE approach is that it produces consistent estimates of regression parameters even if the working correlation structure is misspecified. However, this consistency property can only be obtained when the covariates are not time-dependent or the mean, given all past, present and future values, is equal to the marginal mean. If this assumption does not hold, highly biased estimates might be produced (Pepe and Anderson, 1994).

Let y_{it} be the outcome for individual i at time t, $i = 1, \ldots, N$, $t = 1, \ldots, d$ and let $x_{it} = (x_{it1}, \ldots, x_{itp})'$ be the corresponding covariate vector. We are interested in a marginal model for the mean

$$\mu_{it} = \mathrm{E}(y_{it}|x_{it}) = g^{-1}(x_{it}'\beta), \quad t = 1, \ldots, d \tag{6.1.1}$$

where g is the link function and β is a p dimensional regression parameter of interest.

Let β_0 be an estimator such that $U(\beta_0) = 0$, where $U(\beta) = 0$ is the GEE equation. Then under the assumption that the estimating equation is asymptotically unbiased in the sense that

$$\lim_{N \to \infty} \mathrm{E}_{\beta_0}[U(\beta_0)] = 0 \tag{6.1.2}$$

and suitable regularity conditions, $\hat{\beta}_{GEE} = \beta_0$ is consistent no matter the correlation matrix of y_i is correctly specified or not.

For time-dependent covariates, the assumption (6.1.2) may not hold. A sufficient condition for the GEE estimate $\hat{\beta}_{GEE}$ to be consistent is that the marginal mean is equal to the fully-conditional mean (Pepe and Anderson, 1994)

$$\mathrm{E}(y_{it}|x_{it}) = \mathrm{E}(y_{it}|x_{i1}, \ldots, x_{id}), \quad \text{for } t = 1, \ldots, d. \tag{6.1.3}$$

Pan, Louis and Connett (2000) provided analytical calculation for bias of mean parameter estimates in an autoregressive process model:

$$y_{it}|(y_{i,t-1}, x_{it}) = y_{i,t-1} + x_{it}\beta + e_{it},$$

where $y_{i0} = 0$, x_{it} are independently and identically distributed (iid) from a normal distribution $N(0, \sigma^2)$, e_{it} are iid from $N(0, \tau^2)$, and x_{it} and e_{it} are independent of each other and of $y_{i,t-1}$, $t = 1, \ldots, d$. β is a scalar. It is easy to verify that the assumption (6.1.3) does not hold and biased GEE estimate is produced.

Assumption (6.1.3) is also a sufficient condition that

$$\mathrm{E}_{\beta_0}\left[\frac{\partial \mu_{is}}{\partial \beta_j}\{y_{it} - \mu_{it}(\beta_0)\}\right] = 0 \quad \text{for all } s, t, \quad s = 1, \ldots, d, \ t = 1, \ldots, d. \tag{6.1.4}$$

Lai and Small (2007) classified time-dependent covariates into three types – types I, II and III. Let $x^j, j = 1, \ldots, p$ denote the jth covariate. A time-dependent covariate x^j is of type I if condition (6.1.4) holds. A time-dependent covariate x^j is of **type II** if it satisfies

$$\mathrm{E}_{\beta_0}\left[\frac{\partial \mu_{is}}{\partial \beta_j}\{y_{it} - \mu_{it}(\beta_0)\}\right] = 0 \quad \text{for all } s \geq t, \quad t = 1, \ldots, d. \tag{6.1.5}$$

Note that the class of type I covariates is a subset of the class of type II covariates. A time-dependent covariate is of type III if it is not of type II, i.e.

$$\mathrm{E}_{\beta_0}\left[\frac{\partial \mu_{is}}{\partial \beta_j}\{y_{it} - \mu_{it}(\beta_0)\}\right] \neq 0 \quad \text{for some } s > t. \tag{6.1.6}$$

The assumption (6.1.1) about the marginal model guarantees that

$$\mathrm{E}_{\beta_0}\left[\frac{\partial \mu_{it}}{\partial \beta_j}\{y_{it} - \mu_{it}(\beta_0)\}\right] = 0, \quad t = 1, \ldots, d, \ j = 1, \ldots, p. \tag{6.1.7}$$

However, the GEE approach uses equation (2.4.1) to combine the d^2p equations (6.1.4) for $j = 1, \ldots, p$. Because some of these estimating equations are not valid for type II and type III, use of an arbitrary working correlation structure may produce inconsistent estimates. But GEEs with independent working correlations combine only the estimating equations (6.1.7) and thus are consistent as long as assumption (6.1.1) holds (see Lai and Small 2007 for more detailed discussions).

Lai and Small (2007) used the generalized method of moments to make optimal use of the estimating equations that are available by the covariates. The GEE approach to use weighted estimating equations might be possible. However, the weight matrix may not be a covariance matrix any more. In the following, we consider the case of

type II covariates only and show that the GEE approach is possible to combine all valid estimating equations.

Denote the inverse of the weight matrix W_i^{-1} in GEE equations by $[h_{st}]$. Then $\left(\frac{\partial \mu_i}{\partial \beta}\right)^T W_i^{-1}(y_i - \mu_i)$ can be expressed as $\left\{\sum_{s=1}^{d}\sum_{t=1}^{d} \frac{\partial \mu_{is}}{\partial \beta_j} h_{st}(y_{it} - \mu_{it})\right\}_{j=1}^{p} = 0$. A sufficient condition for a time-dependent covariate to be type II is

$$\mathrm{E}(y_{it}|x_{it}) = \mathrm{E}(y_{it}|x_{it}, x_{i,t+1} \dots, x_{id}), \quad \text{for } t = 1, \dots, d. \tag{6.1.8}$$

That is, the mean given the current does not depend on the future covariates. This model is reasonable in practice. Based on this condition, the valid estimating equations are given by (6.1.5). If we use the GEE approach to combine all these estimating equations (6.1.5), then in GEE equations the weight matrix $H_i = W_i^{-1}$ should be a lower-triangular matrix and so is W_i. That is,

$$U(\beta) = \sum_{i=1}^{N}\left[\sum_{t=1}^{d}\sum_{s\geq t}^{d} \frac{\partial \mu_{is}}{\partial \beta_j} H_{i[st]}(y_{it} - \mu_{it})\right] = 0, \quad \text{for } j = 1, \dots, p. \tag{6.1.9}$$

Note that the weight matrix H_i is no longer a covariance matrix because of its lower-triangular form. The question is how do we choose the weight matrix. One option is to use Cholesky decomposition of a covariance matrix. Some limited simulations have been done. The proposed approach does not work well all the time. Other choices of weight matrix should be studied.

Appendix A: Derivation of $\frac{\partial^2 l}{\partial \beta^2}$

Using equation (3.2.2) it can be seen that the (k, k')th element of the matrix $\frac{\partial^2 l}{\partial \beta^2}$ can be written as

$$\begin{aligned}
\frac{\partial^2 l}{\partial \beta_k \partial \beta_{k'}} &= \sum_{i=1}^{N} \frac{\partial}{\partial \beta_k} \left\{ \left(\frac{\partial \mu_i}{\partial \beta_{k'}} \right)^T W_i^{-1} (y_i - \mu_i) \right\} \\
&+ \frac{1}{2} \frac{\partial}{\partial \beta_k} \sum_{i=1}^{N} \operatorname{tr} \left[\left\{ W_i^{-1} (y_i - \mu_i)(y_i - \mu_i)^T - I_d \right\} W_i^{-1} \frac{\partial W_i}{\partial \beta_{k'}} \right] \\
&= \operatorname{tr} \sum_{i=1}^{N} W_i^{-1} \left[(y_i - \mu_i) \left(\frac{\partial^2 \mu_i}{\partial \beta_k \partial \beta_{k'}} \right)^T - \frac{\partial W_i}{\partial \beta_k} W_i^{-1} (y_i - \mu_i) \left(\frac{\partial \mu_i}{\partial \beta_{k'}} \right)^T - \frac{\partial W_i}{\partial \beta_k} \left(\frac{\partial \mu_i}{\partial \beta_{k'}} \right)^T \right] \\
&+ \frac{1}{2} \operatorname{tr} \sum_{i=1}^{N} \left\{ \left[\frac{\partial W_i^{-1}}{\partial \beta_k} (y_i - \mu_i)(y_i - \mu_i)^T - W_i^{-1} \frac{\partial \mu_i}{\partial \beta_k} (y_i - \mu_i)^T - W_i^{-1} (y_i - \mu_i) \left(\frac{\partial \mu_i}{\partial \beta_k} \right)^T \right] \right. \\
&\left. W_i^{-1} \frac{\partial W_i}{\partial \beta_{k'}} + \left[W_i^{-1} (y_i - \mu_i)(y_i - \mu_i)^T - I_d \right] \left(\frac{\partial W_i^{-1}}{\partial \beta_k} \frac{\partial W_i}{\partial \beta_{k'}} + W_i^{-1} \frac{\partial^2 W_i}{\partial \beta_k \partial \beta_{k'}} \right) \right\},
\end{aligned} \tag{A.1}$$

where $\frac{\partial^2 W_i}{\partial \beta_{k'} \partial \beta_k} = \frac{\partial^2 A_i^{\frac{1}{2}}}{\partial \beta_{k'} \partial \beta_k} R(\rho) A_i^{\frac{1}{2}} + \frac{\partial A_i^{\frac{1}{2}}}{\partial \beta_{k'}} R(\rho) \frac{\partial A_i^{\frac{1}{2}}}{\partial \beta_k} + \frac{\partial A_i^{\frac{1}{2}}}{\partial \beta_k} R(\rho) \frac{\partial A_i^{\frac{1}{2}}}{\partial \beta_{k'}} + A_i^{\frac{1}{2}} R(\rho) \frac{\partial^2 A_i^{\frac{1}{2}}}{\partial \beta_{k'} \partial \beta_k}$, $k, k' = 1, \ldots, p$.

Appendix B: Proof of asymptotic unbiasedness of equation (3.2.3)

From equation (3.2.2) it can be seen that $\mathrm{E}\left(\frac{\partial l}{\partial \beta_k}\right) = \frac{1}{2}\,\mathrm{tr}\sum_{i=1}^{N}\left\{W_i^{-1}\Sigma_i - I_d\right\}W_i^{-1}\frac{\partial W_i}{\partial \beta_k}$, where $\Sigma_i = \mathrm{Cov}(y_i)$. Now suppose the estimate $R(\hat{\rho})$ of the working correlation converges to the true correlation matrix $C(\rho)$ in probability. Then, asymptotically, as $N \to \infty$, $W_i^{-1}\Sigma_i = A_i^{-1/2}R^{-1}(\hat{\rho})A_i^{-1/2}\ A_i^{1/2}C(\rho)A_i^{1/2} = A_i^{1/2}R^{-1}(\hat{\rho})C(\rho)A_i^{-1/2} = I_d$. Thus, $\mathrm{E}\left(\frac{\partial l}{\partial \beta}\right) = \sum_{i=1}^{N}\mathrm{E}\left(\frac{\partial l_i}{\partial \beta}\right) = 0$, so that the estimating equations (3.2.3) are asymptotically unbiased.

Appendix C: Proof of Theorem 3.2.1

Suppose that $\hat{\beta}$ is consistent. Then the estimate of the correlation parameter ρ_{tu} is given by $\hat{\rho}_{tu} = \sum_{i=1}^{N} y_{it}^{*} y_{iu}^{*} / N, t, u = 1, \ldots, d, t \neq u$. Now, we consider the four cases as what follows.

Case 1: The true correlation structure $C(\rho)$ is unstructured. Now, $\mathrm{E}(y_{it}^{*} y_{iu}^{*}) = \rho_{tu}$. Then, as $N \to \infty$, $\hat{\rho}_{tu}$ converges in probability to ρ_{tu}.

Case 2: The true correlation structure is the general autocorrelation matrix $R(\rho_1, \ldots, \rho_{d-1})$. Then for each $t \neq u$, $\mathrm{E}(y_{it}^{*} y_{iu}^{*}) = \rho_{|t-u|}$, $t, u = 1, \ldots, d$, $i = 1, \ldots, N$. Then, as $N \to \infty$, $\hat{\rho}_{tu}$ converges in probability to $\rho_{|t-u|}$.

Case 3: The true correlation structure is the exchangeable correlation structure $C(\rho)$ in which the diagonal elements are 1 and the off-diagonal elements are ρ. Let c_{tu} be the (t, u) element of $C(\rho)$, $t \neq u$. Under the exchangeable structure, for each $t \neq u$, $c_{tu} = \rho$ and $\mathrm{E}(y_{it}^{*} y_{iu}^{*}) = \rho$, $i = 1, \ldots, N$. Then, as $N \to \infty$, $\hat{\rho}_{tu}$ converges in probability to ρ.

Case 4: The true correlation structure is the AR(1) correlation structure $C(\rho)$ in which the diagonal elements are 1 and the off-diagonal elements are $\rho^{|t-u|}, t \neq u$. Let c_{tu} be the (t, u) element of $C(\rho)$, $t \neq u$. Under the AR(1) structure, for each $t \neq u$, $c_{tu} = \rho^{|t-u|}$ and $\mathrm{E}(y_{it}^{*} y_{iu}^{*}) = \rho^{|t-u|}$, $i = 1, \ldots, N$. Then, as $N \to \infty$, $\hat{\rho}_{tu}$ converges in probability to $\rho^{|t-u|}$.

Therefore, given a consistent estimate of β, the moment estimate of the unstructured working correlation matrix converges in probability to the true correlation matrix irrespective of whether the true correlation structure is unstructured, general autocorrelation, exchangeable or AR(1).

Appendix D: Expressions for $\mathrm{E}\left\{\frac{\partial^2 l_i}{\partial\beta_k\partial\beta_{k'}}\right\}$, $\mathrm{Var}(\frac{\partial l_i}{\partial\beta_k})$ and $\mathrm{Cov}(\frac{\partial l_i}{\partial\beta_k}, \frac{\partial l_i}{\partial\beta_{k'}})$

By taking expectation of the right hand side of equation (A.1), it can be easily seen that

$$\mathrm{E}\left(\frac{\partial^2 l}{\partial\beta_k\partial\beta_{k'}}\right) = -\,\mathrm{tr}\sum_{i=1}^{N} W_i^{-1}\frac{\partial\mu_i}{\partial\beta_{k'}}\left(\frac{\partial\mu_i}{\partial\beta_k}\right)^T$$
$$+\frac{1}{2}\,\mathrm{tr}\sum_{i=1}^{N}\left\{\frac{\partial W_i^{-1}}{\partial\beta_{k'}}\Sigma_i W_i^{-1}\frac{\partial W_i}{\partial\beta_k} + \left(W_i^{-1}\Sigma_i - I_d\right)\left(\frac{\partial W_i^{-1}}{\partial\beta_{k'}}\frac{\partial W_i}{\partial\beta_k} + W_i^{-1}\frac{\partial^2 W_i}{\partial\beta_{k'}\partial\beta_k}\right)\right\}.$$

Now, from equation (3.2.2) we see that

$$\begin{aligned}
\mathrm{Var}\left(\frac{\partial l_i}{\partial\beta_k}\right) &= \mathrm{Var}\left\{\left(\frac{\partial\mu_i}{\partial\beta_k}\right)^T W_i^{-1}(y_i-\mu_i)\right\} + \frac{1}{4}\,\mathrm{Var}\left\{(y_i-\mu_i)^T\frac{\partial W_i^{-1}}{\partial\beta_k}(y_i-\mu_i)\right\}\\
&\quad - \mathrm{Cov}\left\{\left(\frac{\partial\mu_i}{\partial\beta_k}\right)^T W_i^{-1}(y_i-\mu_i), (y_i-\mu_i)^T\frac{\partial W_i^{-1}}{\partial\beta_k}(y_i-\mu_i)\right\}\\
&= \left(\frac{\partial\mu_i}{\partial\beta_k}\right)^T W_i^{-1}\Sigma_i W_i^{-1}\frac{\partial\mu_i}{\partial\beta_k}\\
&\quad + \frac{1}{4}\left[\mathrm{E}\left\{(y_i-\mu_i)^T\frac{\partial W_i^{-1}}{\partial\beta_k}(y_i-\mu_i)(y_i-\mu_i)^T\frac{\partial W_i^{-1}}{\partial\beta_k}(y_i-\mu_i)\right\}\right.\\
&\quad \left. - \left\{\mathrm{E}\left((y_i-\mu_i)^T\frac{\partial W_i^{-1}}{\partial\beta_k}(y_i-\mu_i)\right)\right\}^2\right]\\
&\quad - \mathrm{E}\left\{\left(\frac{\partial\mu_i}{\partial\beta_k}\right)^T W_i^{-1}(y_i-\mu_i)(y_i-\mu_i)^T\frac{\partial W_i^{-1}}{\partial\beta_k}(y_i-\mu_i)\right\}.
\end{aligned}$$

Then, using the expected value of a quadratic form $\mathrm{E}(X^TAX) = \mathrm{tr}(AV) + \mu^TA\mu$, where X is a random vector such that $\mu = \mathrm{E}(X)$ and $V = \mathrm{Var}(X)$ and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, where AB and BA are square matrices, we obtain

$$\begin{aligned}
&\mathrm{Var}\left(\frac{\partial l_i}{\partial\beta_k}\right) = \left(\frac{\partial\mu_i}{\partial\beta_k}\right)^T W_i^{-1}\Sigma_i W_i^{-1}\frac{\partial\mu_i}{\partial\beta_k}\\
&+\frac{1}{4}\left[\mathrm{E}\left\{(y_i-\mu_i)^T\frac{\partial W_i^{-1}}{\partial\beta_k}(y_i-\mu_i)(y_i-\mu_i)^T\frac{\partial W_i^{-1}}{\partial\beta_k}(y_i-\mu_i)\right\} -\right.\\
&\left.\left\{\mathrm{tr}\left(\frac{\partial W_i^{-1}}{\partial\beta_k}\Sigma_i\right)\right\}^2\right] - \mathrm{E}\left\{\left(\frac{\partial\mu_i}{\partial\beta_k}\right)^T W_i^{-1}(y_i-\mu_i)(y_i-\mu_i)^T\frac{\partial W_i^{-1}}{\partial\beta_k}(y_i-\mu_i)\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{\partial \mu_i}{\partial \beta_k}\right)^T W_i^{-1}\Sigma_i W_i^{-1}\frac{\partial \mu_i}{\partial \beta_k} \\
&\quad + \frac{1}{4}\left[\operatorname{tr}\mathrm{E}\left\{(y_i-\mu_i)(y_i-\mu_i)^T\frac{\partial W_i^{-1}}{\partial \beta_k}(y_i-\mu_i)(y_i-\mu_i)^T\frac{\partial W_i^{-1}}{\partial \beta_k}\right\} - \right. \\
&\quad \left.\left\{\operatorname{tr}\left(\frac{\partial W_i^{-1}}{\partial \beta_k}\Sigma_i\right)\right\}^2\right] - \operatorname{tr}\mathrm{E}\left\{(y_i-\mu_i)\left(\frac{\partial \mu_i}{\partial \beta_k}\right)^T W_i^{-1}(y_i-\mu_i)(y_i-\mu_i)^T\frac{\partial W_i^{-1}}{\partial \beta_k}\right\}.
\end{aligned}
$$

Further, by the trace property $\operatorname{tr}(ABCD) = (\mathrm{vec}D)^T(A\otimes C^T)\mathrm{vec}(B^T)$, where A, B, C and D are four matrices such that the matrix product $ABCD$ is defined and square, we obtain

$$
\begin{aligned}
&\mathrm{Var}\left(\frac{\partial l_i}{\partial \beta_k}\right) = \left(\frac{\partial \mu_i}{\partial \beta_k}\right)^T W_i^{-1}\Sigma_i W_i^{-1}\frac{\partial \mu_i}{\partial \beta_k} \\
&+ \frac{1}{4}\left[\mathrm{vec}^T\left(\frac{\partial W_i^{-1}}{\partial \beta_k}\right)\mathrm{E}\left\{(y_i-\mu_i)(y_i-\mu_i)^T\otimes(y_i-\mu_i)(y_i-\mu_i)^T\right\}\mathrm{vec}\left(\frac{\partial W_i^{-1}}{\partial \beta_k}\right)\right. \\
&\left. -\left\{\operatorname{tr}\left(\frac{\partial W_i^{-1}}{\partial \beta_k}\Sigma_i\right)\right\}^2\right] \\
&- \mathrm{vec}^T\left(\frac{\partial W_i^{-1}}{\partial \beta_k}\right)\mathrm{E}\left\{(y_i-\mu_i)\otimes(y_i-\mu_i)(y_i-\mu_i)^T\right\}\mathrm{vec}\left(W_i^{-1}\frac{\partial \mu_i}{\partial \beta_k}\right).
\end{aligned}
$$

By similar calculations, and again, by using the identities $\mathrm{E}(X^TAX) = \operatorname{tr}(AV) + \mu^TA\mu$, $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ and $\operatorname{tr}(ABCD) = (\mathrm{vec}D)^T(A\otimes C^T)\mathrm{vec}(B^T)$, and simplification, it can be shown that

$$
\begin{aligned}
&\mathrm{Cov}\left(\frac{\partial l_i}{\partial \beta_k}, \frac{\partial l_i}{\partial \beta_{k'}}\right) = \left(\frac{\partial \mu_i}{\partial \beta_k}\right)^T W_i^{-1}\Sigma_i W_i^{-1}\frac{\partial \mu_i}{\partial \beta_{k'}} \\
&- \frac{1}{2}\left\{\mathrm{vec}^T\left(\frac{\partial W_i^{-1}}{\partial \beta_k}\right)\mathrm{E}[(y_i-\mu_i)\otimes(y_i-\mu_i)(y_i-\mu_i)^T]\mathrm{vec}\left(W_i^{-1}\frac{\partial \mu_i}{\partial \beta_{k'}}\right)\right\} \\
&- \frac{1}{2}\left\{\mathrm{vec}^T\left(\frac{\partial W_i^{-1}}{\partial \beta_{k'}}\right)\mathrm{E}[(y_i-\mu_i)\otimes(y_i-\mu_i)(y_i-\mu_i)^T]\mathrm{vec}\left(W_i^{-1}\frac{\partial \mu_i}{\partial \beta_k}\right)\right\} \\
&+ \frac{1}{4}\left\{\mathrm{vec}^T\left(\frac{\partial W_i^{-1}}{\partial \beta_{k'}}\right)\mathrm{E}\left[(y_i-\mu_i)(y_i-\mu_i)^T\otimes(y_i-\mu_i)(y_i-\mu_i)^T\right]\mathrm{vec}\left(\frac{\partial W_i^{-1}}{\partial \beta_k}\right)\right. \\
&\left. - \operatorname{tr}\left(\frac{\partial W_i^{-1}}{\partial \beta_k}\Sigma_i\right)\operatorname{tr}\left(\frac{\partial W_i^{-1}}{\partial \beta_{k'}}\Sigma_i\right)\right\}.
\end{aligned}
$$

As can be seen, these expressions for $\mathrm{E}\left\{\frac{\partial^2 l_i}{\partial\beta_k \partial\beta_{k'}}\right\}$, $\mathrm{Var}(\frac{\partial l_i}{\partial\beta_k})$ and $\mathrm{Cov}(\frac{\partial l_i}{\partial\beta_k}, \frac{\partial l_i}{\partial\beta_{k'}})$ require second, third and fourth order simple moments, such as, $\mathrm{E}(y_{iq} - \mu_{iq})^3$ and mixed moments, such as $\mathrm{E}(y_{iq} - \mu_{iq})^2(y_{ir} - \mu_{ir})$, of binary data. These are given in Appendix E.

Appendix E: High order moments of y_i's

Denote $V_i = \mathrm{E}(y_i y_i^T) = \Sigma_i + \mu_i \mu_i^T$. Then, noting that for the binary random variable y, the distribution of each of the random variables y^2, y^3 and y^4 is the same as the distribution of y and by some simple algebra, it can be shown that for $q, r, s = 1, \ldots, d$

$$\mathrm{E}(y_{iq} - \mu_{iq})^2 = \mu_{iq}(1 - \mu_{iq}),$$

$$\mathrm{E}(y_{iq} - \mu_{iq})^3 = \mu_{iq} - 3\mu_{iq}^2 + 2\mu_{iq}^3,$$

$$\mathrm{E}(y_{iq} - \mu_{iq})^4 = -3\mu_{iq}^4 + 6\mu_{iq}^3 - 4\mu_{iq}^2 + \mu_{iq},$$

$$\mathrm{E}(y_{iq} - \mu_{iq})^2(y_{ir} - \mu_{ir}) = (1 - 2\mu_{iq})[V_i(q,r) - \mu_{iq}\mu_{ir}],$$

$$\mathrm{E}(y_{iq} - \mu_{iq})^3(y_{ir} - \mu_{ir}) = (1 - 3\mu_{iq} + 3\mu_{iq}^2)[V_i(q,r) - \mu_{iq}\mu_{ir}],$$

$$\begin{aligned}\mathrm{E}(y_{iq} - \mu_{iq})^2(y_{ir} - \mu_{ir})^2 &= (1 - 2\mu_{iq})(1 - 2\mu_{ir})V_i(q,r) + (1 - 2\mu_{iq})\mu_{iq}\mu_{ir}^2 \\ &\quad + (1 - 2\mu_{ir})\mu_{iq}^2\mu_{ir} + \mu_{iq}^2\mu_{ir}^2,\end{aligned}$$

$$\begin{aligned}&\mathrm{E}(y_{iq} - \mu_{iq})^2(y_{ir} - \mu_{ir})(y_{is} - \mu_{is}) = \\ &\quad (1 - 2\mu_{iq})[\mathrm{E}(y_{iq}y_{ir}y_{is}) - V_i(q,r)\mu_{is} - V_i(q,s)\mu_{ir} + \mu_{iq}\mu_{ir}\mu_{is}] + \mu_{iq}^2\Sigma_i(r,s),\end{aligned}$$

where $V_i(q,r)$ is the (q,r)th element in matrix V_i.

We still need to evaluate $\mathrm{E}(y_{iq}y_{ir}y_{is})$, $\mathrm{E}(y_{iq} - \mu_{iq})(y_{ir} - \mu_{ir})(y_{is} - \mu_{is})$ and $\mathrm{E}(y_{iq} - \mu_{iq})(y_{ir} - \mu_{ir})(y_{is} - \mu_{is})(y_{it} - \mu_{it})$ for $q, r, s, t = 1, \ldots, d$. These quantities cannot be obtained for binary data. So we approximate these by using results from the multivariate normal distribution which are given by

$$E(y_{iq}y_{ir}y_{is}) = \mu_{iq}\Sigma_i(r,s) + \mu_{ir}\Sigma_i(q,s) + \mu_{is}\Sigma_i(q,r) + \mu_{iq}\mu_{ir}\mu_{is},$$

$$E(y_{iq} - \mu_{iq})(y_{ir} - \mu_{ir})(y_{is} - \mu_{is}) = 0,$$

and

$$E(y_{iq} - \mu_{iq})(y_{ir} - \mu_{ir})(y_{is} - \mu_{is})(y_{it} - \mu_{it}) =$$

$$\Sigma_i(q,r)\Sigma_i(s,t) + \Sigma_i(q,s)\Sigma_i(r,t) + \Sigma_i(q,t)\Sigma_i(r,s).$$

Appendix F: Derivation of κ_{ij}, $\kappa_{ij}^{(l)}$ and κ_{ijl}

As mentioned earlier we treat the generalized estimating function (4.2.2) as if it were a likelihood score function. By the decoupling method of Crowder (2001) where the working covariance matrix is regarded as a constant matrix with respect to the regression parameters β, the first derivative by using the chain rule and the product rule in matrix calculus (Magnus and Neudecker, 1988) of $U(\beta;\rho,\phi)$ with respect to β is

$$\frac{\partial U}{\partial \beta} = \sum_{n=1}^{N}\left[(X_n^T \otimes y_n^T W_n^{-1})\frac{\partial \Delta_n}{\partial \mu_n} - (\Delta_n X_n)^T W_n^{-1} - (X_n^T \otimes \mu_n^T W_n^{-1})\frac{\partial \Delta_n}{\partial \mu_n}\right]\Delta_n X_n,$$

where $\frac{\partial \Delta_n}{\partial \mu_n}$ is a $d^2 \times d$ dimensional sparse matrix with non-zero quantities $f'(F^{-1}(\mu_{nj}))$ $(F^{-1})'(\mu_{nj})$ in the $[(j-1)d+j, j]$ term, $j = 1,\ldots,d$, $n = 1,\ldots,N$.

It is easy to see that

$$I = \{-\kappa_{ij}\} = -\mathrm{E}\left(\frac{\partial U(\beta;\rho,\phi)}{\partial \beta}\right) = \sum_{n=1}^{N}(\Delta_n X_n)^T W_n^{-1}\Delta_n X_n \tag{F.1}$$

and

$$\begin{aligned}\left(\{\kappa_{ij}^{(1)}\},\{\kappa_{ij}^{(2)}\},\cdots,\{\kappa_{ij}^{(p)}\}\right)^T &= \frac{\partial}{\partial \beta}\left\{\mathrm{E}\left(\frac{\partial U(\beta;\rho,\phi)}{\partial \beta}\right)\right\}\\ &= -\sum_{n=1}^{N}(X_n^T \otimes X_n^T)\left[(\Delta_n W_n^{-1})\otimes I_d + I_d \otimes (\Delta_n W_n^{-1})\right]\frac{\partial \Delta_n}{\partial \mu_n},\end{aligned} \tag{F.2}$$

where I_d is a d-dimensional identity matrix. Further, the second derivative of U by using the chain rule, the product rule and the Kronecker product rule in matrix calculus with respect to β is

$$\begin{aligned}
\frac{\partial^2 U}{\partial \beta^2} &= \sum_{n=1}^{N} \Bigg\{ \frac{\partial}{\partial \mu_n} \left[(X_n^T \otimes y_n^T W_n^{-1}) \frac{\partial \Delta_n}{\partial \mu_n} \Delta_n X_n \right] - \frac{\partial}{\partial \mu_n} \left[(\Delta_n X_n)^T W_n^{-1} \Delta_n X_n \right] \\
&\quad - \frac{\partial}{\partial \mu_n} \left[(X_n^T \otimes \mu_n^T W_n^{-1}) \frac{\partial \Delta_n}{\partial \mu_n} \Delta_n X_n \right] \Bigg\} \Delta_n X_n \\
&= \sum_{n=1}^{N} \Bigg\{ (I_p \otimes X_n^T \otimes y_n^T W_n^{-1}) \left[(X_n^T \Delta_n) \otimes I_{d^2} \cdot \frac{\partial^2 \Delta_n}{\partial \mu_n^2} + (X_n^T \otimes \frac{\partial \Delta_n}{\partial \mu_n}) \frac{\partial \Delta_n}{\partial \mu_n} \right] \\
&\quad - (X_n^T \otimes X_n^T) \left[(\Delta_n W_n^{-1}) \otimes I_d + I_d \otimes (\Delta_n W_n^{-1}) \right] \frac{\partial \Delta_n}{\partial \mu_n} \\
&\quad - \left[(\frac{\partial \Delta_n}{\partial \mu_n} \Delta_n X_n)^T \otimes I_p \right] (I_d \otimes K_{dp}) (\mathrm{vec}(X_n^T) \otimes I_d) \cdot W_n^{-1} \\
&\quad - (I_p \otimes X_n^T \otimes \mu_n^T W_n^{-1}) \left[((X_n^T \Delta_n) \otimes I_{d^2}) \frac{\partial^2 \Delta_n}{\partial \mu_n^2} + (X_n^T \otimes \frac{\partial \Delta_n}{\partial \mu_n}) \frac{\partial \Delta_n}{\partial \mu_n} \right] \Bigg\} \Delta_n X_n,
\end{aligned}$$

where K_{dp} is a $dp \times dp$ commutation matrix and $\frac{\partial^2 \Delta_n}{\partial \mu_n^2}$ is a $d^3 \times d$ dimensional sparse matrix with non-zero quantities $f''(F^{-1}(\mu_{nj})) \left[(F^{-1})'(\mu_{nj}) \right]^2 + f'(F^{-1}(\mu_{nj}))(F^{-1})''(\mu_{nj})$ in the $[d(d+1)(j-1)+j, j]$ term, $j = 1, \ldots, d$, $n = 1, \ldots, N$. Then, after a few steps of algebra, we obtain

$$\begin{aligned}
(\{\kappa_{ij1}\}, \{\kappa_{ij2}\}, \cdots, \{\kappa_{ijp}\})^T &= \mathrm{E}\left(\frac{\partial^2 U(\beta; \rho, \phi)}{\partial \beta^2} \right) \\
&= - \sum_{n=1}^{N} \Bigg\{ (X_n^T \otimes X_n^T) \left[(\Delta_n W_n^{-1}) \otimes I_d + I_d \otimes (\Delta_n W_n^{-1}) \right] \frac{\partial \Delta_n}{\partial \mu_n} \\
&\quad + \left[(\frac{\partial \Delta_n}{\partial \mu_n} \Delta_n X_n)^T \otimes I_p \right] (I_d \otimes K_{dp}) (\mathrm{vec}(X_n^T) \otimes I_d) \cdot W_n^{-1} \Bigg\} \Delta_n X_n.
\end{aligned} \tag{F.3}$$

Bibliography

[1] Ashford, J. R. and Sowden, R. R. (1970). Multivariate probit analysis. *Biometrics*, **26**, 535-46.

[2] Carey, V., Zeger, S. L. and Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, **80**, 517-26.

[3] Chaganty, N. R. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations. *Journal of Statistical Planning and Inference*, **63**, 39-54.

[4] Chaganty, N. R. and Joe, H. (2004). Efficiency of generalized estimating equations for binary responses. *ournal of the Royal Statistical Society, Series B*, **66**, 851-60.

[5] Chaganty, N. R. and Shults, J. (1999). On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter. *Journal of Statistical Planning and Inference*, **76**, 127-44.

[6] Cordeiro, G. M. and Klein, R. (1994). Bias correction in ARMA models. *Statistics and Probability Letters*, **19**, 169-76.

[7] Cox, D. R. and Snell, E. J. (1968). A general definition of residuals (with discussion), *Journal of the Royal Statistical Society, Series B*, **30**, 248-75.

[8] Crowder, M. (1985). Gaussian estimation for correlated binomial data. *Journal of the Royal Statistical Society, Series B*, **47**, 229-37.

[9] Crowder, M. (1986). On consistency and inconsistency of estimating equations. *Economic Theory*, **2**, 305-30.

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

[11] Crowder, M. (2001). On repeated measures analysis with misspecified covariance structure. *Journal of the Royal Statistical Society, Series B*, **63**, 55-62.

[12] Davidian, M. and Carroll, R. J. (1987) Variance function estimation. *Journal of the American Statistical Association*, **82**, 1079-91.

[13] Davidian, M. and Giltinan, D. M. (1995). *N*onlinear Models For Repeated Measurement Data. Chapman and Hall: London.

[14] Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data*, Oxford University Press: Oxford.

[15] Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27-38.

[16] Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**, 141-51.

[17] Fitzmaurice, G. M., Laird, N. M. and Lipsitz, S. R. (1994). Analysing incomplete longitudinal binary responses: a likelihood-based approach. *Biometrics*, **50**, 601-12.

[18] Fitzmaurice, G. M., Laird, N. M. and Rotnitzky, A. G. (1993). Regression models for discrete longitudinal responses. *Statistical Science*, **8**, 284-99.

[19] Godambe, V. P. (1960). An optimum property of regular maximum likelihood equation. *The Annals of Mathematical Statistics*, **31**, 1208-11.

[20] Hand, D. and Crowder, M. (1996). *Practical Longitudinal Data Analysis.* Chapman & HALL: London.

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

[22] Huggins, R. M. (1993). A robust approach to the analysis of repeated measures. *Biometrics*, **49**, 715-20.

[23] Jørgenson, B. (1997). *T*he Theory of Dispersion Modles, Chapman & Hall: London.

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

[25] Lai, T. L. and Small, D. (2007). Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method of moments approach. *Journal of the Royal Statistical Society, Series B*, **69**, 79-99.

[26] Laird, N. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963-74.

[27] Liang, K.-Y. and Zeger S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 45-51.

[28] Liang, K.-Y. and Zeger S. L. (1995). Inference based on estimating functions in the presence of nuisance parameters. *Statistical Science*, **10**, 158-73.

[29] Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: using odds ratios as a measure of association. *Biometrika* **78**, 153-60.

[30] Lipsitz, S. R, Fitzmaurice, G. M., Sleeper, L. and Zhao, L. P. (1995). Estimation methods for the joint distribution of repeated binary observations. *Biometrics*, **51**, 562-70.

[31] Magnus, J. R. and Neudecker, H. (1988). *M*atrix Differential Calculus with Applications in Statistics and Econometrics, John Wiley & Sons: Chichester.

[32] McCullagh, P. (1983). Quasi-Likelihood functions. *Annals of Statistics*, **11**, 59-67.

[33] McCullagh, P. and Nelder, J. A. (1983) *Generalized Linear Models.* Chapman & HALL: London.

[34] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370-84.

[35] Pan, W., Louis, T. A. and Connett, J. E. (2000). A note on marginal linear regression with correlated response data. *The American Statistician*, **54**, 191-95.

[36] Paul, S. R. and Banerjee, T. (1998). Analysis of two-way layout of count data involving multiple counts in each cell. *Journal of the American Statistical Association*, **93**, 1419-29.

[37] Paul, S. R. and Islam, A. S. (1998). Joint estimation of the mean and dispersion parameters in the analysis of proportions: a comparison of efficiency and bias. *Canadian Journal of Statistics*, **26**, 83-94.

[38] Paul, S. R. and Plackett, R. L. (1978). Inference sensitivity for Poisson mixtures. *Biometrika*, **65**, 591-602.

[39] 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 Computation*, **23**, 939-51.

[40] Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**, 1033-48.

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

[42] Qin, G. Y., Zhu, Z. Y. and Fung, W. K. (2008). Robust estimating equations and bias correction of correlation parameters for longitudinal data. *Computational Statistics and Data Analysis*, **52**, 4745-53.

[43] Qu, A., Lindsay, B. G. and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, **87**, 823-36.

[44] Rao, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, **52**, 447-58.

[45] Rosner, B. (1984). Multivariate methods in ophtalmology with applications to other paired data situations. *Biometrics*, **40**, 1025-35.

[46] Shults, J. and Chaganty, N. R., (1998). Analysis of serially correlated data using quasi-least squares. *Biometrics*, **54**, 1622-30.

[47] Song, P. X.-K. (2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics*, **27**, 305-20.

[48] Song, P. X.-K., Li, M. and Yuan Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics*, **65**, 60-68.

[49] Stefanescu, C. and Turnbull, B. W. (2005). On the multivariate probit model for exchangeable binary data with covariates. *Biometrical Journal*, **47**, 206-18.

[50] Stiratelli, R. Laird, N. and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, **40**, 961-71.

[51] Sutradhar, B. C. (2003). An overview on regression models for discrete longitudinal responses. *Statistical Science*, **18**, 377-93.

[52] Sutradhar, B. C. and Das, K. (1999). On the efficiency of regression estimators in generalised linear models for longitudinal data. *Biometrika*, **86**, 459-65.

[53] Sutradhar, B. C. and Das, K. (2000). On the accuracy of efficiency of estimating equation approach. *Biometrics*, **56**, 622-25.

[54] Wang, Y.-G. and Carey, V. (2003). Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance. *Biometrika*, **90**, 29-41.

[55] Wang, Y.-G. and Carey, V. J. (2004). Unbiased estimating equations from working correlation models for irregularly timed repeated measures. *Journal of the American Statistical Association*, **99**, 845-53.

[56] Wang, Y.-G. and Lin, X. (2005). Effects of variance-function misspecification in anlysis of longitudinal data. *Biometrics*, **61**, 413-21.

[57] Wang, Y.-G., Lin, X. and Zhu, M. (2005). Robust estimating functions and bias correction for longitudinal data analysis. *Biometrics*, **61**, 684-91.

[58] Wang, Y.-G. and Zhao, Y. (2007). A modified pseudolikelihood approach for analysis of longitudinal data. *Biometrics*, **63**, 681-89.

[59] Wedderburn, R. W. M. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, **61**, 439-47.

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

[61] Woolson, R. F. and Clarke, W. R. (1984). Analysis of categorical incomplete longitudinal data. *Journal of the Royal Statistical Society, Series A*, **147**, 87-99.

[62] Yahav, I. and Shmueli, G. (2011). On generating multivariate Poisson data in management science applications. *Applied Stochastic Models in Business and Industry*, doi: 10.1002/asmb.901.

[63] Ye, H. and Pan J. (2006). Modelling of covariance structures in generalised estimating equations for longitudinal data. *Biometrika*, **93**, 927-41.

[64] Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121-30.

[65] Zeger, S. L. and Qaqish, B. (1988). Markov regressiion models for time series: a quasi-likelihood approach. *Biometrics*, **44**, 1049-60.

Vita Auctoris

Mr. Xuemao Zhang was born in 1976 in Shandong, China. He got his Bachelor degree in Mathematics Education from Qufu Normal University. In 2003 he came to University of Windsor to continue his graduate study. He earned his Master degree in Pure Mathematics in 2005 and Master degree in Statistics in 2007. He completed his Ph.D. degree in Statistics in 2011.