GENERALIZED ESTIMATING EQUATIONS FOR MIXED MODELS

Lulah Alnaji

A Dissertation

Submitted to the Graduate College of Bowling Green State University in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2018

Committee:

Hanfeng Chen, Advisor

Robert Dyer, Graduate Faculty Representative

Wei Ning

Junfeng Shang

Copyright ©August 2018 Lulah Alnaji All rights reserved

ABSTRACT

Hanfeng Chen, Advisor

Most statistical approaches of molding the relationship between the explanatory variables and the responses assume subjects are independent. However, in clinical studies the longitudinal data are quite common. In this type of data, each subject is assessed repeatedly over a period of time. Therefore, the independence assumption is unlikely to be valid with longitudinal data due to the correlated observations of each subject. Generalized estimating equations method is a popular choice for longitudinal studies. It is an efficient method since it takes the within-subjects correlation into account by introducing the $n \times n$ working correlation matrix $R(\alpha)$ which is fully characterized by the correlation parameter α . Although the generalized estimating equations' methodology considers correlation among the repeated observations on the same subject, it ignores the between-subject correlation and assumes subjects are independent.

The objective of this dissertation is to provide an extension to the generalized estimating equations to take both within-subject and between-subject correlations into account by incorporating the random effect b to the model. If our interest focuses on the regression coefficients, we regard the correlation parameter α and as nuisance and estimate the fixed effects β using the estimating equations $U(\beta, \hat{G}, \hat{\alpha})$. If our interest focuses either on both β and the variance of the random effects b or on the coefficient parameters and the association structure, then building an additional system of estimating equations analogous to $U(\beta, G, \alpha)$ can serve to estimate either β and G, simultaneously or β and α , simultaneously. In this later two cases the correlation matrix must be specified carefully. It is sensitive to the misspecification of the working correlation matrix $R(\alpha)$ in contrast to the first case which allows to the misspecification of the working correlation matrix $R(\alpha)$ when we are interested on the fixed effects parameter only. Moreover, the later two cases require the first four moments to be specified while the first case depends only on the first two moments. This estimating equations method has no closed form solution and can be solved iteratively. For example, Newton-Raphson is a popular iterative method to be used. We illustrate through simulation studies and real data applications the performance of the proposed methods in terms of bias and efficiency. Moreover, we investigate their behaviors compared to those for existing methods such as generalized estimating equations (GEE), generalized linear models (GLM) and generalized linear mixed models (GLMM). For further studying the performance of newly proposed method, the new approach is applied to the epilepsy data that was studied by many others Fitzmaurice, Laird, and Ware (2012). For my beloved parents, siblings, my loving husband Hassan and my gorgeous children Judy Amjad, Reematy and Meelad. It is an amazing feeling to finish working on my dissertation and words alone cannot say how happy that I am to spend evenings with you and give back the time that I spent working on this dissertation. I love you tremendously and to you I dedicate this work.

ACKNOWLEDGMENTS

First of all, my sincerest gratitude and respect to my advisor, Dr. Hanfeng Chen, for his continuous support, suggestions and advice throughout the course of this dissertation. Completion of my dissertation would not have been possible without his support and guidance. He inspired me to be an independent researcher and was always there to support and advice me.

My gratitude also goes to my other committee members: Dr. Robert Dyer, Dr. Wei Ning, and Dr. Junfeng Shang for their valuable time and advice. My sincere thanks goes to Dr. Craig Zirble for his online LATEX resource which helped me in the implementation of this dissertation. I would like to extend my gratitude to all my professors in the Department of Mathematics and Statistics at BGSU who taught and mentored me. I am very thankful to our department staff Anna Lynch, Carol Nungester and Amber Snyder for their help.

I am forever grateful to the Saudi Arabian government for their generous financial assistantship during the five years at Bowling Green State University, without their financial support I would not even have been able to start my graduate school, yet alone finish it. I am very fortunate to have had the opportunity to study in United States of America, to everyone who made this possible I cannot thank you enough.

I am very thankful to my parents and siblings for their unlimited support, love and for their faith in me. Thank you for encouraging me, making me smile and being always a great source of support. I would like to thank my friends Yi-Ching Lee, Amani Alghamdi, Rajha Alghamdi and my neighbor Sue Wade who made my stay in Bowling Green more supported, memorable and happier. I am very grateful to anyone who has walked alongside me and guided me during my study life.

I am thankful to my loving and understanding husband Hassan and my gorgeous children Judy Amjad, Reematy and Meelad who make my life meaningful for the patience and understanding for the time that I spent away from them throughout my graduate school.

Lulah Alnaji

Bowling Green, OH

TABLE OF CONTENTS

			Page
CHAPT	ER 1	LITERATURE REVIEW	. 1
1.1	Introdu	uction	. 1
1.2	Genera	alized estimating equations	. 3
	1.2.1	Link function	. 3
	1.2.2	Variance function	. 4
	1.2.3	Working correlation matrix	. 5
	1.2.4	GEE	. 8
1.3	Second	d-order of generalized estimating equations (GEE2)	. 11
1.4	Linear	mixed-effects models	. 14
1.5	Generalized linear mixed model		
1.6	Structu	are of the Dissertation	. 17
СНАРТ	ER 2	GENERALIZED ESTIMATING EQUATIONS FOR MIXED MODELS .	. 19
2.1	2.1 Introduction		
2.2	Incorporating random effects in GEE		
	2.2.1	Generalized estimating equations for mixed model	. 20
	2.2.2	Estimating α and G	. 21
	2.2.3	Estimation Parameter β	. 21
2.3	GEEM	I for count longitudinal data	. 24
	2.3.1	Marginal means, variances and covariances	. 25
2.4	Simula	ation study	. 25
	2.4.1	Simulation 1	. 26

			viii
	2.4.2	Simulation 2	31
	2.4.3	Plots for simulated data when $K = 50$	31
	2.4.4	Plots for simulated data when $K = 100$	35
	2.4.5	Plots for simulated data when $K = 150$	38
CHAPT	ER 3 S	SECOND-ORDER GENERALIZED ESTIMATING EQUATIONS FOR MIXE	ED
MOI	DELS .		46
3.1	GEE2	for mixed models	46
	3.1.1	Estimation β and G , simultaneously	47
	3.1.2	Estimating β and α , simultaneously	52
3.2	Simula	tion study	54
CHAPT	ER 4 1	REAL DATA APPLICATIONS	60
4.1	Introdu	ction	60
4.2	Data de	escription	60
4.3	A GEE	M Model for the Seizure Data	68
4.4	A GEI	EM2 Model for Epilepsy Data	76
CHAPT	ER 5 \$	SUMMARY AND CONCLUSION	78
5.1	Conclu	sion Remarks	78
5.2	Future	Research	80
BIBLIO	GRAPH	ΙΥ	81
APPENI	DIX A	SELECTED R PROGRAMS	86

LIST OF FIGURES

Figure	I	Page
2.1	Boxplots of numbers of observations when $K = 50$	32
2.2	Boxplots of log of numbers of observations for $K = 50$	32
2.3	Trajectories of observations of each cluster in the sample when $K = 50. \ldots$	33
2.4	Trajectories of observations of each cluster in the sample when $K = 50$ together.	34
2.5	Boxplots of numbers of observations when $K = 100.$	35
2.6	Boxplots of log of numbers of observations for $K = 100.$	35
2.7	Trajectories of observations of each cluster in the sample when $K = 100.$	36
2.8	Trajectories of observations of each cluster in the sample when $K = 100.$	37
2.9	Boxplots of numbers of observations when $K = 150.$	38
2.10	Boxplots of log of numbers of observations for $K = 150$	38
2.11	Trajectories of observations of each cluster in the sample when $K = 150.$	39
2.12	Trajectories of observations of each cluster in the sample when $K = 150.$	40
4.1	Number of seizures per each subject in the sample during 8-week prior to the treat-	
4.2	ment. . <td>61</td>	61
	to the treatment.	62
4.3	Number of seizures per each subject in Progabide group over 8-week period prior	
	to the treatment.	62
4.4	Boxplots of log numbers of seizures for the placebo group during the baseline	
	period and during visits post randomization.	64

4.5	Boxplots of numbers of seizures for the Progabide group during the baseline period	
	and during visits post randomization	64
4.6	Boxplots of log of numbers of seizures for the placebo group during the baseline	
	period and during visits post randomization.	65
4.7	Boxplots of log of numbers of seizures for the Progabide group during the baseline	
	period and during visits post randomization.	65
4.8	Number of seizures per each subject in the sample during 8-week prior to the treat-	
	ment	66
4.9	Number of seizures per each subject in the sample during 8-week prior to the treat-	
	ment	67
4.10	The plot of the random effects of random intercept and random coefficient Time	
	for placebo group.	74
4.11	Resorting the plot of the random effects of random intercept and random coefficient	
	Time for placebo group.	74
4.12	The plot of the random effects of random intercept and random coefficient Time	
	for Progabide group.	75
4.13	Resorting the random effects of random intercept and random coefficient Time for	
	Progabide group.	75

Х

LIST OF TABLES

Table	Page
1.1	Link and variance functions of some distributions of exponential family McCullagh
	and Nelder (1989)
1.2	The most common choices of the working correlation matrix where $N = \sum n_i$
	Molenberghs and Verbeke (2005)
2.1	Parameter estimates (standard deviations in parentheses) of the true beta values
	$(3.00, 0.50, 1.00, 0.20)$ when number of clusters $K = 50, 100$ and 150 with $n_i =$
	4 per each cluster, is compared for: the newly proposed method (GEEM), GEE
	Liang and Zeger (1986), GLM McCullagh (1984) and GLMM method, for $5,000$
	simulations applying the model 2.4.1 with correlation matrix as shown in 2.4.13 41
2.2	Parameter estimates (standard deviations in parentheses) of the true beta values
	$(3.00, 0.50, 1.00, 0.20)$ when number of clusters $K = 50, 100$ and 150 with $n_i =$
	4 per each cluster, is compared for: the newly proposed method (GEEM), GEE
	Liang and Zeger (1986), GLM McCullagh (1984) and GLMM method, for $5,000$
	simulations applying the model 2.4.1 with correlation matrix as shown in 2.4.10 42
2.3	Parameter estimates (standard deviations in parentheses) of the true beta values
	$(3.00, 0.50, 1.00, 0.20)$ when number of clusters $K = 50, 100$ and 150 with $n_i =$
	4 per each cluster, is compared for: the newly proposed method (GEEM), GEE
	Liang and Zeger (1986), GLM McCullagh (1984) and GLMM method, for $5,000$
	simulations applying the model 2.4.1 with correlation matrix as shown in 2.4.9 43

- 2.4 Parameter estimates (standard deviations in parentheses) of the true beta values (3.00, 0.50, 1.00, 0.20) when number of clusters K = 50, 100 and 150 with $n_i =$ 4 per each cluster, is compared for: the newly proposed method (GEEM), GEE Liang and Zeger (1986), GLM McCullagh (1984) and GLMM method, for 5,000 simulations applying the model 2.4.1 with correlation matrix as shown in 2.4.11 . . . 44

4.1	The data of some participants in epilepsy study. The data was downloaded from	
	website: "http://www.hsph.harvard.edu/fitzmaur/ala2e/epilepsy.sas7bdat"	69
4.2	Parameter estimates for fixed effects part applying the newly proposed GEEM ap-	
	proach on epilepsy data	72
4.3	Parameter estimates for applying the generalized estimating equation method Liang	
	and Zeger (1986) using geeglm() function in R language	73
4.4	Parameter estimates for applying the generalized estimating equation method Liang	
	and Zeger (1986) using gee () function in R language	73
4.5	Parameter estimates for fixed effects part and variance of random effect applying	
	the proposed GEEM2 approach on epilepsy data.	76
4.6	Parameter estimates for fixed effects part and the correlation parameter applying	
	the proposed GEEM2 approach on epilepsy data.	77

xiii

CHAPTER 1 LITERATURE REVIEW

1.1 Introduction

The essential job for many clinical studies is to model the relationship between the explanatory variables and the response using statistical models such as general linear model and generalized linear model. The point is to estimate the parameters of interest included in the model and interpret the results. The response follows either a continuous or discrete distribution but the observations are assumed to be independent, so it can be handled by such models.

Following up and collecting data repeatedly from the same subject such as blood pressure, cholesterol level, etc are quite common in medical studies. With the increasing availability of longitudinal data, the assumption of independent observations is not ideal. Taking the correlation between observations (within-subject correlation) into account increases the efficiency of regression parameter estimation. Numerous models have been proposed for this purpose. These models can be divided into two major categories referred to as conditional and marginal models.

Generalized Linear Model (GLM) proposed by Nelder and Baker (1972) is a common framework to estimate the regression coefficients of linear models. The word "Generalized" points to non-Gaussian distributions since GLM methodology is applicable to any distribution belongs to the exponential family distributions either continuous or discrete. This approach was extended by McCullagh and Nelder (1983) and McCullagh (1984) to accommodate the longitudinal data with the assumption that the repeated observations per subject are independent. These approaches are helpful, however, assuming that the repeated measurements are independent for the same subject while in fact they are correlated may effect the efficiency of these approaches.

To overcome these issues, Liang and Zeger (1986) and Zeger and Liang (1986) extended these approaches to Generalized Estimating Equations (GEE) by introducing a working correlation matrix $R_i(\alpha)$ which is an $n_i \times n_i$ a correlation matrix and fully characterizes the matrix V_i . The matrix V_i is the analogous matrix to the variance-covariance matrix of Y_i . GEE became a popular methodology to estimate the regression parameters of marginal distributions for correlated when the correlation is regarded as a nuisance. Generalized estimating equations is known to provide consistent estimators even if the working correlation matrix is not accurate.

GEE methodology was extended to GEE1 by Prentice (1988) for binary outcomes by devolving an additional estimating equation for the association parameters. This approach can model the regression coefficients β and the correlation parameter α , simultaneously. Later on, GEE1 was extended to GEE2 by Prentice and Zhao (1991) for discrete and continuous responses in the exponential family. Unlike GEE, GEE1 and GEE2 approaches are sensitive to the misspecification of the working correlation matrix as illustrated by Heagerty and Zeger (1996)

Subjects are often assumed to be independent and no between-subject correlation. However, patients that go to the same clinic have correlated data. The within-subject correlation is often taken into account in generalized estimating equations (GEE) by the correlation matrix introduced by Liang and Zeger (1986).

In this dissertation, we extend the GEE approach into a more general case to incorporate both within-subject and between-subject associations. We consider three scenarios: (i) considering the regression coefficient β are the parameters of interest. (ii) considering both the regression coefficient and the association structure are the parameters of interest. (iii) considering both the regression coefficient and the variance of the random effects are the parameters of interest.

1.2 Generalized estimating equations

Let $Y_i = (y_{i1}, y_{i2}, \dots, y_{in_i})'$ be the $n_i \times 1$ response vector generated from a distribution in the exponential family, $X_i = (x_{i1}, x_{i2}, \dots, x_{in_i})'$ is the $n_i \times p$ vector of covariates for the *i*-th subject $i = 1, \dots, K$ corresponding to the fixed effects $\beta \in \mathbb{R}^p$, and ϵ_i is the model error.

In matrix notation,

$$Y_{i} = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_{i}} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_{0} \\ \beta_{1} \\ \vdots \\ \beta_{p} \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in_{i}} \end{bmatrix}$$
$$X_{i} = \begin{bmatrix} x'_{i1} \\ x'_{i2} \\ \vdots \\ \vdots \\ x'_{in_{i}} \end{bmatrix} = \begin{bmatrix} x_{i11} & x_{i12} & x_{i13} & \dots & x_{i1p} \\ x_{i21} & x_{i22} & x_{i23} & \dots & x_{i2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{in_{i}1} & x_{in_{i}2} & x_{in_{i}3} & \dots & x_{in_{ip}} \end{bmatrix}$$

The generalized estimating equation methodology is applicable to any distribution belonging to the exponential family. Therefore, the response can be continuous or discrete. The maximization of the likelihood of the longitudinal data is extremely difficult. GEE provides consistent estimates depending only on the first two moments

$$g(\mathbf{E}(y_{ij})) = x'_{ij}\beta$$

$$v(\mathbf{E}(Y_i)) = v(\mu_i)$$
(1.2.1)

where function g called the link function, it connects $E(y_{ij})$ to $x'_{ij}\beta$, μ_i is a vector consists of $\mu_{ij} = E(y_{ij})$ and $v(\cdot)$ is known variance function (See Table 1.1) Liang and Zeger (1986). 1.2.1 Link function

Modeling data set using generalized estimating equations methodology requires determining three essential components. *Random Component* which specifies the probability distribution of the

response variable from the exponential family. Systematic Component (explanatory variables) that particularizes the explanatory variables $(x_{i1}, x_{i2}, \dots, x_{in_i})$ in the model. These variables are linear in parameters $(\beta_1, \dots, \beta_p)$, the linear combination $(x'_{ij}\beta)$ called linear predictor Hutcheson and Sofroniou (1999). The third component is the Link Function $g(\cdot)$ which connects the previous two components (i.e the random and the systematic components) by relating the expected value of the y_{ij} to the random component as

$$g(\mathbf{E}(y_{ij})) = x'_{ij}\beta$$
 (1.2.2)

The appropriate $g(\cdot)$ is the function that makes the relationship between the transformed mean and the systematic component linear. The link function is assumed to be monotonically increasing in μ_i which guarantees each value of $X'_i\beta$ has only one corresponding to $E(Y_i) = \mu_i$ and differentiable to ensure that the coefficient parameters β can be estimated Swan (2006).

In the normal distribution, the mean and the predictor parameters range from $-\infty$ to ∞ and that preforms a linear relationship, and hence the link function $g(\cdot)$ is identity. Since in GEE the distribution of the response can be any distribution from the exponential family. So, it can be continuous or discrete which in some distributions the relationship between the predictor and the expected value of Y_i is not linear and requires link function to be used to address this issue. For instance, the mean of the binary distribution ranges from 0 to 1 but the predicted parameter ranges from $-\infty$ to ∞ . An appropriate link function can transform the mean of the binary distribution from [0, 1] to $(-\infty, \infty)$ McCullagh and Nelder (1989). (See Table 1.1 for the common choices of link functions for distribution such as Normal, Poisson, Binomial, Gamma and Inverse Gamma)

1.2.2 Variance function

The variance function $v(\mu_i)$ characterizes the variance as it depends on the mean. In other words, it expresses the relationship between the mean and the variance.

The variance function is meaningful since it allows the mean and the variance to be communicate in a unique way, and hence it classifies the members in the class of exponential family distribution Firth (1991). The links and variance functions that are commonly used for Normal, Poisson, Binomial, Gamma and Inverse-Gamma distributions are summarized by McCullagh and Nelder (1989) as in Table 1.1

Distribution	Notation	Link	Variance Function
Normal	$N(\mu, \sigma^2)$	Identity	$v(\mu) = 1$
Poisson	$P(\lambda)$	log	$v(\mu) = \lambda$
Binomial	$B(m,\pi)/m$	logit	$v(\mu) = \mu(1-\mu)$
Gamma	$G(\mu, v)$	reciprocal	$v(\mu) = \mu^2$
Inverse Gamma	$IG(\mu, \sigma^2)$	$1/\mu$	$v(\mu) = \mu^3$

Table 1.1: Link and variance functions of some distributions of exponential family McCullagh and Nelder (1989)

Generalized estimating equations methodology by Liang and Zeger (1986) is in fact an extension to Generalized Linear Model (GLM) proposed by McCullagh and Nelder (1983). GLM is an extensive treatment that refers to a wide class of models. In this model, $y_{i1}, \dots, y_{in_i}, i = 1, \dots, K$ are assumed to be independent random variables, and drawn from a distribution in the exponential family.

1.2.3 Working correlation matrix

The working correlation matrix $R_i(\alpha)$ is an $n_i \times n_i$ matrix that is fully specified by the unknown correlation parameter α . The main rule of this correlation matrix is assuming the within-subject association is known since the within-subject correlation is rarely known. With the existence of the matrix $R_i(\alpha)$ the coefficient parameters can be estimated based on pairwise correlated responses which improve the efficiency of the standard errors.

Unfortunately, there is no determined way of choosing a specific working correlation matrix, it is completely left to the researchers' own discretion and this is one of the complications of the GEE approach Swan (2006). Pankhurst, Connolly, Jones, and Dobson (2003) recommended choosing $R_i(\alpha)$ carefully by ensuring it is consistent with the empirical correlations. The standard choice structures of the working correlation matrix are independent, unstructured, exchangeable, and autoregressive (AR(1)) Zeger and Liang (1986).

(1) (Independent structure) This is the basic form of the working correlation matrix forms. The independent structure is basically the identity matrix $(R_i = I_{ni})$ and has no α to be estimated, since it assumes no pairwise within-subject association,

$$\operatorname{Corr}(y_{ij}, y_{ik}) = 0, \ \forall \ j \neq k$$

This form is unlikely to be valid for longitudinal data since the repeated observations for the same subject are highly correlated. Using this type of structure can lead to large loss in efficiency. This structure is used in GLM which is helpful to find an initial β estimate to be used in GEE algorithm. The function of the independent working correlation is defined as

$$R_{i,j} = \begin{cases} 1 & i = j \\ 0 & \text{otherwise.} \end{cases}$$

In matrix notation,

$$R_i = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

(2) (*Unstructured structure*) This type of working correlation matrix has no specific structure as no constraints are imposed to values of this correlation matrix.

$$\operatorname{Corr}(y_{ij}, y_{ik}) = \alpha_{jk} = \alpha_{kj} = \operatorname{Corr}(y_{ik}, y_{ij}), \ \forall \ j \neq k$$

It assumes that all pairwise associations are different. This form is easy to understand but at the same time it is very computationally expensive to be estimated, especially in the large data set as it requires the contrivance to be estimated for each pair of times (i.e. all $n_i(n_i - 1)/K$). It can be written as follows

$$R_{i,j} = \begin{cases} 1 & i = j \\ \\ \alpha_{ij} & \text{otherwise.} \end{cases}$$

Or in matrix notation,

$$R_{i} = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \dots & \alpha_{1n_{i}} \\ \alpha_{21} & 1 & \alpha_{23} & \dots & \alpha_{2n_{i}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{n_{i}1} & \alpha_{n_{i}2} & \alpha_{n_{i}3} & \dots & 1 \end{bmatrix}$$

(3) (*Exchangeable structure*) This form is also called *compound symmetry*. It assumes the observations within a subject are equally correlated

$$\operatorname{Corr}(y_{ij}, y_{ik}) = \alpha, \ \forall \ j \neq k$$

The big feature of this correlation matrix is that, only one parameter needs to be estimated but it ignores the time varying between observations. It can be written as

$$R_{i,j} = \begin{cases} 1 & i = j \\ \alpha & \text{otherwise.} \end{cases}$$

Or in matrix notation,

$$R_{i} = \begin{bmatrix} 1 & \alpha & \alpha & \dots & \alpha \\ \alpha & 1 & \alpha & \dots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \dots & 1 \end{bmatrix}$$

(4) (Autoregressive structure) First Order Autoregressive AR(1) assumes the correlation in-

creases as if the measurements are closer in time and decreases as the distance getting farther between time points.

$$\operatorname{Corr}(y_{ij}, y_{ik}) = \alpha^{|j-k|}, \ \forall \ j \neq k$$

The autoregressive structure is defined as

$$R_{i,j} = \begin{cases} 1 & i = j \\ \alpha^{|j-k|} & \text{otherwise.} \end{cases}$$

In matrix notation,

$$R_{i} = \begin{bmatrix} 1 & \alpha^{|j-k|} & \alpha^{|j-k|} & \dots & \alpha^{|j-k|} \\ \\ \alpha^{|j-k|} & 1 & \alpha^{|j-k|} & \dots & \alpha^{|j-k|} \\ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \\ \alpha^{|j-k|} & \alpha^{|j-k|} & \alpha^{|j-k|} & \dots & 1 \end{bmatrix}$$

These examples of the working correlation structures are summarized in Table 1.2

1.2.4 GEE

The assumption that the pairwise correlation among the repeated measures (α) is not zero

$$\operatorname{Corr}(y_{ij}, y_{ik}) = \alpha_{ijk}$$

extends GLM method to GEE Liang and Zeger (1986) and Zeger and Liang (1986). The exact covariance matrix of Y_i is unknown, but if the working correlation matrix $R(\alpha)$ is chosen correctly then the approximation of the covariance matrix of Y_i is

$$V_{i_{\text{GEE}}} = A^{1/2} R_i(\alpha) A^{1/2}$$
(1.2.3)

In other words,

$$Cov(Y_i) \approx A_i^{1/2} R_i(\alpha) A_i^{1/2}$$
 (1.2.4)

where $R_i(\alpha)$ is the working correlation matrix which fully characterizes by α and A_i is $n_i \times n_i$ diagonal matrix consists of the second moment of Y_i

$$A_i = \operatorname{diag}\{v(\mu_i)\}$$

The coefficient parameter vector β is obtained by solving the estimating equations,

$$U_{\text{GEE}}(\beta, \alpha(\beta)) = \sum_{i=1}^{K} \left(D_{i_{\text{GEE}}}(\beta, \alpha(\beta)) \right)' \left(V_{i_{\text{GEE}}}(\beta, \alpha(\beta)) \right)^{-1} \left(S_{i_{\text{GEE}}}(\beta, \alpha(\beta)) \right) = 0, \quad (1.2.5)$$

where $S_{i_{\text{GEE}}} = (Y_i - \mu_i)$, $D_{i_{\text{GEE}}} = \partial(\mathbf{E}(Y_i))/\partial\beta$ and $V_{i_{\text{GEE}}} = A_i^{1/2} R_i A_i^{1/2}$ is as defined in 1.2.3 which is a function of β and α . As a result, (1.2.5) is a function of β and α . Changing (1.2.5) to be a function of β only, can be done by replacing α by its estimates $\hat{\alpha}$ Liang and Zeger (1986),

$$U_{\text{GEE}}(\beta, \hat{\alpha}(\beta)) = \sum_{i=1}^{K} \left(D_{i_{\text{GEE}}}(\beta, \hat{\alpha}(\beta)) \right)' \left(V_{i_{\text{GEE}}}(\beta, \hat{\alpha}(\beta)) \right)^{-1} \left(S_{i_{\text{GEE}}}(\beta, \hat{\alpha}(\beta)) \right) = 0, \quad (1.2.6)$$

and $\hat{\beta}$ is a solution to the (1.2.6) which is used as an estimate for β . To fit this model, a popular method of finding a solution of GEE is solving the resulting (nonlinear) equations iteratively using the Fisher "method of scoring" algorithm.

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + \left[\sum_{i=1}^{K} \left(D_{i_{\text{GEE}}}(\beta, \hat{\alpha}(\beta)) \right)' \left(V_{i_{\text{GEE}}}(\beta, \hat{\alpha}(\beta)) \right)^{-1} D_{i_{\text{GEE}}}(\beta, \hat{\alpha}(\beta)) \right]^{-1} \times \left[\sum_{i=1}^{K} \left(D_{i_{\text{GEE}}}(\beta, \hat{\alpha}(\beta)) \right)' \left(V_{i_{\text{GEE}}}(\beta, \hat{\alpha}(\beta)) \right)^{-1} S_{i_{\text{GEE}}}(\beta) \right]$$
(1.2.7)

In general, the parameter α can be estimated from Pearson residuals

$$\hat{r}_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}$$
(1.2.8)

using the method of moments. Table 1.2 presents by Molenberghs and Verbeke (2005) shows the

common choices of the working covariance matrix.

Structure	$\operatorname{Corr}(Y_{ij}, Y_{ik})$	Estimator
Independence	0	-
Exchangeable	α	$\hat{\alpha} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i(n_i-1)} \sum_{i \neq j} r_{ij} r_{ik}$
AR(1)	$\alpha^{ j-k }$	$\hat{\alpha} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i - 1} \sum_{i \le n_i - 1} r_{ij} r_{i,j+1}$
Unstructured	$lpha_{ jk }$	$\hat{\alpha} = \frac{1}{N} \sum_{i=1}^{N} r_{ij} r_{ik}$

Table 1.2: The most common choices of the working correlation matrix where $N = \sum n_i$ Molenberghs and Verbeke (2005)

In this approach the estimates $\hat{\beta}$ are still valid (i.e. consistent) even if one misspecified the correlation structure with loss of efficiency but the standard errors may not Crowder (2001). Moreover, the parameter estimates $\hat{\beta}$ is asymptotically normally distributed

$$V_{\text{GEE}} = \lim_{K \to \infty} K \left(\sum_{i=1}^{K} D'_{i_{\text{GEE}}} V_{i_{\text{GEE}}}^{-1} D_{i_{\text{GEE}}} \right)^{-1} \left(\sum_{i=1}^{K} D'_{i_{\text{GEE}}} V_{i_{1}}^{-1} Cov(Y_{i}) V_{i_{\text{GEE}}}^{-1} D_{i_{\text{GEE}}} \right) \times \left(\sum_{i=1}^{K} D'_{i_{\text{GEE}}} V_{i_{\text{GEE}}}^{-1} D_{i_{\text{GEE}}} \right)^{-1}$$
(1.2.9)

as $K \to \infty$ with zero mean.

The data differs from one another and hence the patterns of correlation between observations may vary among studies. Gaining more efficiency in standard errors requires specifying the pairwise correlation pattern correctly.

GLM and GEE are very similar, but in some cases the GEE method is chosen over GLM method. In GLM we assume that the repeated observations for a subject are independent. This is very unlikely to be valid if $n_i > 1$ (i.e. repeated observations) and then the correlation must be taken into account. However, if $n_i = 1$ (i.e. single observation) then the GLM can be applied to obtain a description for a variety of continuous or discrete variables y_{ij} .

1.3 Second-order of generalized estimating equations (GEE2)

Zhao and Prentice (1990), Prentice and Zhao (1991), Zhao, Prentice, and Self (1992) and Chaganty (1997) among others have extended the GEE approach in terms of estimating the correlation structure or the matrix $R(\alpha)$ instead of using method of moments to estimate α from the Pearson residual when $R(\alpha) \neq I_{n_i}$. Precisely, Zhao and Prentice (1990) build a system of estimating equations analogous to the system of Liang and Zeger (1986). Let,

$$T_{i} = (Y_{i} - E(Y_{i}))(Y_{i} - E(Y_{i}))'$$

= $(t_{i11}, t_{i22}, \cdots, t_{ijj}, t_{i12}, t_{i23}, \cdots, t_{i,j-1,j})'$ (1.3.1)

be an $n_i(n_i - 1)/2 + n_i \times 1$ vector with $E(T_i) = \zeta_i$. The system below allows to model β and α , simultaneously

$$U_{i_{\text{GEE}}}[\beta, \alpha(\beta)] = \sum_{i=1}^{K} \left(D_{i_{\text{GEE}}}(\beta, \alpha(\beta)) \right)' \left(V_{i_{\text{GEE}}}(\beta, \alpha(\beta)) \right)^{-1} \left(S_{i_{\text{GEE}}}(\beta, \alpha(\beta)) \right) = 0$$

$$U_{i_{\text{GEE2}}}[\beta, \alpha(\beta)] = \sum_{i=1}^{K} \left(D_{i_{\text{GEE2}}}(\beta, \alpha(\beta)) \right)' \left(V_{i_{\text{GEE2}}}(\beta, \alpha(\beta)) \right)^{-1} \left(S_{i_{\text{GEE2}}}(\beta, \alpha(\beta)) \right) = 0$$

$$(1.3.2)$$

where $S_{i_{\text{GEE2}}} = T_i - \zeta_i$, $D_{i_{\text{GEE2}}} = \partial E(T_i)/\partial \alpha$ and $V_{i_{\text{GEE2}}} = \text{Var}(T_i)$. The matrix $V_{i_{\text{GEE2}}}$ looks very similar to the matrix $V_{i_{\text{GEE}}}$ but in fact they are different. $V_{i_{\text{GEE2}}}$ is $n_i(n_i-1)/2 + n_i \times n_i(n_i-1)/2 + n_i$ and since it is not straightforward to determine a working covariance model for T_i , because $\text{var}(T_i)$ requires the third and fourth moments be specified, therefore the independence is often assumed Diggle (2002). The general form of 1.3.2 can be written as

$$U_{GEE2}[\beta, \alpha(\beta)] = \sum_{i=1}^{K} \left(D_{GEE2}(\beta, \alpha(\beta)) \right)^{'} \left(V_{GEE2}(\beta, \alpha(\beta)) \right)^{-1} \left(S_{GEE2}(\beta, \alpha(\beta)) \right)$$
$$= \sum_{i=1}^{K} \left[\begin{array}{c} D_{i11} & D_{i12} \\ D_{i21} & D_{i22} \end{array} \right]^{'} \left[\begin{array}{c} V_{i11} & V_{i12} \\ V_{i21} & V_{i22} \end{array} \right]^{-1} \left[\begin{array}{c} S_{i1} \\ S_{i2} \end{array} \right] = 0$$
(1.3.3)

where,

$$D_{GEE2}(\beta, \alpha(\beta)) = \begin{bmatrix} D_{i11} & D_{i12} \\ D_{i21} & D_{i22} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mu_i}{\partial \beta} & \frac{\partial \mu_i}{\partial \alpha} \\ \frac{\partial \zeta_i}{\partial \beta} & \frac{\partial \zeta_i}{\partial \alpha} \end{bmatrix},$$

$$V_{GEE2}(\beta, \alpha(\beta)) = \begin{bmatrix} V_{i11} & V_{i12} \\ V_{i21} & V_{i22} \end{bmatrix} = \begin{bmatrix} \operatorname{Var}(Y_i) & \operatorname{Cov}(Y_i, T_i) \\ \operatorname{Cov}(T_i, Y_i) & \operatorname{Var}(T_i) \end{bmatrix}$$

$$S_{GEE2}(\beta, \alpha(\beta)) = \begin{bmatrix} S_{i1} \\ S_{i2} \end{bmatrix} = \begin{bmatrix} y_i - \mu_i \\ T_i - \zeta_i \end{bmatrix}$$

However, dealing with the matrix $D_{GEE2}(\beta, \alpha(\beta))$ in 1.3.3 with its actual form raise some potential complications such as the interpretation of a mean vector that contains a correlation parameter is no longer simple. To address this issue assume that the matrix D_i is a diagonal matrix (i.e. $\frac{\partial \mu_i}{\partial \alpha} = 0$ and $\frac{\partial \zeta_i}{\partial \alpha} = 0$) Ziegler, Kastner, and Blettner (1998).

For the working covariance matrix $V_{GEE2}(\beta, \alpha(\beta))$ Prentice and Zhao (1991) consider different structures.

Independent working covariance matrix V_{GEE2}(β, α(β)) by assuming the elements of y_i are independent. Therefore, that the working covariance matrix is diagonal.

$$\operatorname{Cov}(Y_i, T_i) = \operatorname{Cov}(T_i, Y_i) = 0$$

• Gaussian working covariance matrix $V_{GEE2}(\beta, \alpha(\beta))$ by assuming that the elements of y_i are distributed normally, then

$$Cov(y_{ij}, t_{ik}t_{il}) = E((Y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})(y_{il} - \mu_{il})) = 0, \quad \forall \ j, k, l$$

and,

$$\operatorname{Cov}(t_{ij}t_{ik}, t_{il}tim) = \operatorname{E}((y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})(y_{il} - \mu_{il})(y_{im} - \mu_{im})) - \sigma_{ijk}\sigma_{ilm}, \quad \forall \ j, k, l, m \in \operatorname{Cov}(t_{ij}t_{ik}, t_{il}tim) = \operatorname{E}((y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})(y_{il} - \mu_{il})(y_{im} - \mu_{im})) - \sigma_{ijk}\sigma_{ilm}, \quad \forall \ j, k, l, m \in \operatorname{Cov}(t_{ij}t_{ik}, t_{il}tim) = \operatorname{E}((y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})(y_{il} - \mu_{il})(y_{im} - \mu_{im})) - \sigma_{ijk}\sigma_{ilm}, \quad \forall \ j, k, l, m \in \operatorname{Cov}(t_{ij}t_{ik}, t_{il}tim) = \operatorname{E}((y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})(y_{il} - \mu_{il})(y_{im} - \mu_{im})) - \sigma_{ijk}\sigma_{ilm}, \quad \forall \ j, k, l, m \in \operatorname{Cov}(t_{ij}t_{ik}, t_{il}tim) = \operatorname{E}((y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})(y_{il} - \mu_{il})(y_{im} - \mu_{im})) - \sigma_{ijk}\sigma_{ilm}, \quad \forall \ j, k, l, m \in \operatorname{Cov}(t_{ij}t_{ik}) = \operatorname{E}((y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})(y_{il} - \mu_{il})(y_{im} - \mu_{im})) - \sigma_{ijk}\sigma_{ilm}, \quad \forall \ j, k, l, m \in \operatorname{Cov}(t_{ij}t_{ik}) = \operatorname{E}((y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})(y_{ik} - \mu_{ik})(y_{ik} - \mu_{ik})(y_{im} - \mu_{im}))$$

where, σ_{ijk} is the covariance between the *j*-th and *k*-th observations for the *i*-th subject.

 Gaussian working covariance matrix V_{GEE2}(β, α(β)) with common third and fourth correlations. Therefore, in this case we use

$$\mathbf{E}((y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})(y_{il} - \mu_{il})) = \gamma_{jkl} \sqrt{\sigma_{ijj} \sigma_{ikk} \sigma_{ill}}$$

and,

$$E((y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})(y_{il} - \mu_{il})(y_{im} - \mu_{im})) = \sigma_{ijk}\sigma_{ilm} + \sigma_{ijl}\sigma_{ikm} + \sigma_{ijm}\sigma_{ikl} + \delta_{jklm}\sqrt{\sigma_{ijj}\sigma_{ikk}\sigma_{ill}\sigma_{ill}}$$

where the parameter $\{\gamma_{jlm}\}_{j \leq l \leq m}$ and $\{\delta_{jklm}\}_{j \leq k \leq l \leq m}$, can be estimated as

$$\hat{\gamma}_{jlm} = \frac{1}{N} \sum_{i} \frac{((y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})(y_{il} - \mu_{il}))}{\sqrt{\sigma_{ijj}\sigma_{ikk}\sigma_{ill}}}$$

$$\hat{\delta}_{jklm} = \frac{1}{N} \sum_{i} \frac{(y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})(y_{il} - \mu_{il})(y_{im} - \mu_{im}) - \sigma_{ijk}\sigma_{ilm} + \sigma_{ijl}\sigma_{ikm} + \sigma_{ijm}\sigma_{ikl}}{\sqrt{\sigma_{ijj}\sigma_{ikk}\sigma_{ill}\sigma_{ill}}}$$

where N is the total, $N = \sum_{i=1}^{K} n_i$ Lane (2007).

The parameter estimate $\sqrt{K}(\hat{\beta} - \beta), \sqrt{K}(\hat{\alpha} - \alpha)$, is asymptotically normally distributed

$$H = \lim_{K \to \infty} K(H_0 H_1 H_0)$$
(1.3.4)

where

$$H_{0} = \left(\sum_{i=1}^{K} \left(D_{GEE2}(\beta, \alpha(\beta)) \right)^{\prime} \left(V_{GEE2}(\beta, \alpha(\beta)) \right)^{-1} \left(D_{GEE2}(\beta, \alpha(\beta)) \right) \right)^{-1}$$
$$H_{1} = \left(\sum_{i=1}^{K} \left(D_{GEE2}(\beta, \alpha(\beta)) \right)^{\prime} \left(V_{GEE2}(\beta, \alpha(\beta)) \right)^{-1} Cov(Y_{i}) \left(V_{GEE2}(\beta, \alpha(\beta)) \right)^{-1} \left(D_{GEE2}(\beta, \alpha(\beta)) \right) \right)^{-1}$$

as $K \to \infty$ with zero mean.

Since the primary interest lies on β and α , the estimates of β and α are obtained by Newton-Raphson algoritm

$$\begin{bmatrix} \hat{\beta}^{(t+1)} \\ \hat{\alpha}^{(t+1)} \end{bmatrix} = \begin{bmatrix} \hat{\beta}^{(t)} \\ \hat{\alpha}^{(t)} \end{bmatrix} + \left[\sum_{i=1}^{K} D'_{GEE2}(\hat{\beta}^{(t)}, \hat{\alpha}^{(t)}) (V_{GEE2}(\hat{\beta}^{(t)}, \hat{\alpha}^{(t)}))^{-1} D_{GEE2}(\hat{\beta}^{(t)}, \hat{\alpha}^{(t)}) \right]^{-1} \\ \times \left[\sum_{i=1}^{K} D'_{GEE2}(\hat{\beta}^{(t)}, \hat{\alpha}^{(t)}) (V_{GEE2}(\hat{\beta}^{(t)}, \hat{\alpha}^{(t)}))^{-1} S_{GEE2}(\hat{\beta}^{(t)}, \hat{\alpha}^{(t)}) \right]$$
(1.3.5)

Iterate until convergence.

1.4 Linear mixed-effects models

Linear mixed-effects models are extensions of linear regression models for data that are collected and summarized in groups. These models describe the relationship between a response variable and independent variables, with coefficients that can vary with respect to one or more grouping variables. A mixed-effects model consists of two parts, fixed effects and random effects. Fixed-effects terms are usually the conventional linear regression part, and the random effects are associated with individual experimental units drawn at random from a population. The random effects have prior distributions while fixed effects do not. Mixed-effects models can represent the covariance structure related to the grouping of data by associating the common random effects to observations that have the same level of a grouping variable. The classic linear mixed model (LMM) is defined as follows:

$$Y_i = X_i\beta + Z_ib_i + \epsilon_i, \quad i = 1, \cdots, K, \tag{1.4.1}$$

where Y_i is the $n_i \times 1$ response vector, $\beta \in \mathbb{R}^p$ is the fixed effects and $b_i \in \mathbb{R}^q$ is the random effects, X_i is the $n_i \times p$ design matrix corresponding to the fixed effects, Z_i is the $n_i \times q$ design matrix to the random effects, and ϵ_i is the model error. Assume that y_{i1}, \dots, y_{ni} are independent and that ϵ_{ij} and b_i are independent with

$$\epsilon_{ij} \sim N(0, \sigma_{\epsilon}^2 I), \text{ for } i = 1, \cdots, K; j = 1, \cdots, n_i$$

$$b_i \sim N(0, G), \qquad (1.4.2)$$

$$Var(Y_i) = Z'_i G Z_i + \sigma_{\epsilon}^2 I_{ni}$$

Note that the linear model with fixed effect is the special case of $Z_i = 0$. The simple linear mixed model is another special case. Consider

$$y_{ij} = \beta_{i0} + b_{i1}x_{ij} + \epsilon_{ij}, \quad i = 1, \cdots, K, \quad j = 1, \cdots, n_i,$$
(1.4.3)

where y_{ij} is the *j*-th measurement on the *i*-th subject, β_{i0} is the *fixed* intercept parameter and b_{i1} is the random slope parameter for the *i*-th subject, and x_{ij} is a covariate (for observation time or something else).

1.5 Generalized linear mixed model

Although the linear mixed model framework is very useful, its downside is that it is applicable to continuous distributions only. It can not be applied to discrete distributions such as count and binary. The framework of LMM Laird and Ware (1982) has been extended to a more general framework called generalized linear mixed model (GLMM) to accommodate distributions like logistic and log-normal Breslow and Clayton (1993). In fact, the linear mixed model which assumes identity link function is a special case of the generalized linear mixed model.

Let Y_i be the $n_i \times 1$ response vector belongs to a distribution in the exponential family, $\beta \in \mathbb{R}^p$ is the fixed effects and $b_i \in \mathbb{R}^q$ is the random effects, X_i is the $n_i \times p$ design matrix corresponding to the fixed effects, Z_i is the $n_i \times q$ design matrix to the random effects, and ϵ_i is the model error. Assume that y_{i1}, \dots, y_{ini} are independent with the conditional mean and covariance

$$g(\mathbf{E}(y_{ij}|b_i)) = x'_{ij}\beta + x'_{ij}b_i,$$

$$\operatorname{Cov}(Y_i|b_i) = \operatorname{Cov}(\epsilon_i) = \sigma_{\epsilon}^2 I_{ni}$$
(1.5.1)

where g is some known link function. The mutually independent random effects b_i have some probability distribution with zero mean and a covariance matrix G.

$$b_i \sim N(0,G), \ i = 1, \cdots, K.$$
 (1.5.2)

The marginal covariance is given by

$$\operatorname{Cov}(Y_i) = Z'_i G Z_i + \sigma_{\epsilon}^2 I_{ni} \tag{1.5.3}$$

For instance, GLMM can be used to model a count response. Let y_{ij} be generated from a possion distribution with mean λ_{ij} . The common link function for modeling such a response is the

log function (table 1.1)

$$y_{ij} \sim \text{poisson}(\lambda_{ij}), \quad i = 1, \cdots, K, \quad j = 1, \cdots, n_i$$

$$b_i \sim N(0, G), \qquad (1.5.4)$$

$$\log(\mathbf{E}(Y_{ij}|b_i)) = x'_{ij}\beta + z'_{ij}b_i,$$

Patterson and Thompson (1971), Hemmerle and Hartley (1973) and Harville (1977) among others discussed iterative algorithm for Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) in mixed models under the assumption of independent random effects Chen (2010). Later on, Neuhaus and McCulloch (2006) showed independence assumption can result in biased estimates.

1.6 Structure of the Dissertation

The main idea of this dissertation is taking both the pairwise (within-subject) and betweensubject correlation into account aiming to gain more efficiency. In Chapter 2, we first consider the regression coefficients are the parameter of interest. We propose new estimating equations that can accommodate the main idea of this dissertation when the structure of the working correlation matrix is regarded as a nuisance. Based on various working correlation structures (i.e. independent, unstructured, AR(1), MA(1)), with different sample sizes we conduct a simulation study to assess the performance of new approaches in terms of bias and efficiency. After examining the performance of the new approaches, we compare the simulated outcomes with those for some existing elected methods.

In Chapter 3, when either the regression coefficients and correlation structure or the regression coefficients and variance of the random effects are parameters of interest, we extend the approach in Chapter 2 by means of GEE2 to handle all the parameters. Also, we conduct a simulation study to assess the performance of new approaches in terms of bias and efficiency following the simulation studies in the second chapter. We present an iterative method of estimating the parameters. Newton-Raphson algorithm.

CHAPTER 2 GENERALIZED ESTIMATING EQUATIONS FOR MIXED MODELS 2.1 Introduction

In longitudinal studies, the data is collected repeatedly for the same subject over a period of time, occur frequently in clinical trials or medicine. To estimate the regression parameters in a marginal model, it is common to analyze such models using the generalized estimating equations method Liang and Zeger (1986) which requires only the first two moments to be specified. This method is known to provide consistent regression parameter estimates. Generalized estimating equations method can handle the correlation between the pairwise within-subject association but ignores between-subject association. Incorporating random effects in generalized estimating equations.

2.2 Incorporating random effects in GEE

Let $Y_i = (y_{i1}, \dots, y_{in_i})'$ be the $n_i \times 1$ response vector for the *i*-th subject and $X_i = (x_{i1}, \dots, x_{in_i})'$ be the $n_i \times p$ matrix of covariate values for the *i*-th subject $i = 1, \dots, K$ corresponding to the fixed effects $\beta \in R^p$, Z_i is the $n_i \times q$ design matrix to the random effects $b_i \in R^q$, y_{ij} is generated from a distribution in the exponential family with the conditional mean and variance

$$g(\mathbf{E}(Y_i|b_i)) = X'_i\beta + Z'_ib_i$$
(2.2.1)

where g is a link function (See Section 1.2.1). The random effects b_i are assumed to be mutually independent, following the normal distribution with zero mean and covariance matrix G. The exact covariance matrix of Y_i is unknown, but if the working correlation matrix $R(\alpha)$ is chosen correctly then the approximation of the covariance matrix of Y_i is

$$V_{i_{GEEM}} = \text{Cov}[E(Y_i|b_i)] + E[\text{Cov}(Y_i|b_i)]$$

= $Z'_i G Z_i + A_i^{1/2} R_i(\alpha) A_i^{1/2}$ (2.2.2)

In other words,

$$Cov(Y_i) \approx Z'_i G Z_i + A_i^{1/2} R_i(\alpha) A_i^{1/2}$$
 (2.2.3)

where matrix A_i is a diagonal matrix consists of the second moment of the model and $R_i(\alpha)$ is a correlation matrix (See section1.2.3). Therefore, the model error is not necessarily independent. Molenberghs, Verbeke, and Demétrio (2007) use similar idea to 2.2.3, but for Poisson distribution. Moreover, they ignore the association between observation with considering the correlation matrix to be always identity. The model 2.2.3 is more general, it can handle the association between observations and the variation between subjects. Furthermore, our newly proposed model 2.2.3 is applicable to any distribution belonging to the exponential family. While the model proposed by Molenberghs et al. (2007) is only applicable to Poisson distribution. The model proposed by Molenberghs et al. (2007) is studies in this dissertation but with assuming $R_i(\alpha)$ is not identity while they assumed the matrix $R_i(\alpha)$ is identity (See Section 2.3).

2.2.1 Generalized estimating equations for mixed model

Let $\mu_i = (\mu_{i1}, \dots, \mu_{in_i})$ denote the vector of marginal mean of a distribution from exponential family. The generalized estimating equation for mixed model (GEEM) for the coefficient parameters β and the variance of the random effect G is given by

$$U_{\text{GEEM}}[\beta, \alpha(\beta), G] = \sum_{i=1}^{K} \begin{bmatrix} \partial \mu_i(\beta, \alpha, G) / \partial \beta \\ \partial \mu_i(\beta, \alpha, G) / \partial G \end{bmatrix}' \begin{bmatrix} Z_i G Z'_i + A_i^{1/2} R_i(\alpha) A_i^{1/2} \end{bmatrix}^{-1} \\ \times \begin{bmatrix} Y_i - \mu_i(\beta, \alpha, G) \end{bmatrix} \\ = \sum_{i=1}^{K} \left(D_{i_{\text{GEEM}}}(\beta, \alpha, G) \right)' \left(V_{i_{\text{GEEM}}}(\beta, \alpha, G) \right)^{-1} \left(S_{i_{\text{GEEM}}}(\beta, \alpha, G) = 0 \\ (2.2.4) \end{bmatrix}$$

where $V_{i_{\rm GEEM}}$ as defined in 2.2.2 , $S_{i_{\rm GEEM}} = (Y_i - \mu_i(\beta, \alpha, G))$ and

$$D_{i_{\text{GEEM}}} = \begin{bmatrix} \partial \mu_i(\beta, \alpha, G) / \partial \beta \\ \partial \mu_i(\beta, \alpha, G) / \partial G \end{bmatrix}$$

,

Wang, Lee, Zhu, Redline, and Lin (2013). For estimating the coefficient parameters β estimating equations 2.2.4 cannot be used in its actual form due to so many unknown parameters as shown in the next sections.

2.2.2 Estimating α and G

G is the variance matrix for the random effects b_i . Specifically, it is the variance matrix for b_{i0} and b_{i1} . The matrix G can be estimated from the data using Nonlinear Mixed-Effects Models method proposed by Lindstrom and Bates (1990). Generally, α can be estimated as

$$\hat{\alpha}_{uv} = \sum_{i=1}^{K} \frac{\hat{r}_{iu} \hat{r}_{iv}}{N}$$
(2.2.5)

where $N = \sum n_i$ and \hat{r} is the Pearson residual given by 1.2.8 (See Section 1.2).

2.2.3 Estimation Parameter β

It is extremely difficult to estimate β under the actual form of the generalized estimating equation 2.2.4. As a solution, we may employ the estimation of the covariance matrix of the random effects and the estimation of the correlation parameter α . Therefore, the estimation equation of β is given by

$$U_{\text{GEEM}}[\beta, \hat{\alpha}(\beta), \hat{G}] = \sum_{i=1}^{K} \left(\partial \mu_i(\beta, \hat{\alpha}, \hat{G}) / \partial \beta \right)' \left(Z_i \hat{G} Z_i' + A_i^{1/2} R_i(\hat{\alpha}) A_i^{1/2} \right)^{-1} \\ \times \left(Y_i - \mu_i(\beta, \hat{\alpha}, \hat{G}) \right) \\ = \sum_{i=1}^{K} \left(\tilde{D}_{i_{\text{GEEM}}}(\beta, \hat{\alpha}, \hat{G}) \right)' \left(\tilde{V}_{i_{\text{GEEM}}}(\beta, \hat{\alpha}, \hat{G}) \right)^{-1} \left(\tilde{S}_{i_{\text{GEEM}}}(\beta, \hat{\alpha}, \hat{G}) \right) = 0$$

$$(2.2.6)$$

This is the estimating equation 2.2.4 but with replacing the correlation parameter α and the random effects by their estimates. That is,

$$\begin{split} \tilde{S}_{i_{\text{GEEM}}} &= Y_i - \mu_i(\beta, \hat{\alpha}, \hat{G}) \\ \tilde{V}_{i_{\text{GEEM}}} &= Z_i^{'} \hat{G} Z_i + A_i^{1/2} R_i(\hat{\alpha}) A_i^{1/2} \\ \tilde{D}_{i_{\text{GEEM}}} &= \partial \mu_i(\beta, \hat{\alpha}, \hat{G}) / \partial \beta. \end{split}$$

This is nonlinear equations and a popular method of finding a solution of the GEE is solving the resulting (nonlinear) equations iteratively using the Fisher "method of scoring" algorithm. The Fisher scoring iterative equation is

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + \left(\sum_{i=1}^{K} \left(D_{i_{\text{GEEM}}}(\hat{\beta}^{(t)}, \hat{\alpha}, \hat{G}) \right)' \left(\tilde{V}_{i_{\text{GEEM}}}(\hat{\beta}^{(t)}, \hat{\alpha}, \hat{G}) \right)^{-1} \left(D_{i_{\text{GEEM}}}(\hat{\beta}^{(t)}, \hat{\alpha}, \hat{G}) \right) \right)^{-1} \\ \times \left(\sum_{i=1}^{K} \left(D_{i_{\text{GEEM}}}(\hat{\beta}^{(t)}, \hat{\alpha}, \hat{G}) \right)' \left(\tilde{V}_{i_{\text{GEEM}}}(\hat{\beta}^{(t)}, \hat{\alpha}, \hat{G}) \right)^{-1} \left(S_{i_{\text{GEEM}}}(\hat{\beta}^{(t)}, \hat{\alpha}, \hat{G}) \right) \right)$$
(2.2.7)

Theorem 2.2.8. The estimator $\hat{\beta}$ of β is consistent and $\sqrt{K}(\hat{\beta} - \beta)$ has the asymptotic normal distribution with covariance matrix

$$H_{GEEM} = \lim_{K \to \infty} K(H_{0_{GEEM}} H_{1_{GEEM}} H_{0_{GEEM}})$$
(2.2.9)

where

$$H_{0_{GEEM}} = \left(\sum_{i=1}^{K} D'_{i_{GEEM}} V_{i_{GEEM}}^{-1} D_{i_{GEEM}}\right)^{-1}$$
$$H_{0_{GEEM}} = \left(\sum_{i=1}^{K} D'_{i_{GEEM}} V_{i_{GEEM}}^{-1} Cov(Y_i) V_{i_{GEEM}}^{-1} D_{i_{GEEM}}\right)$$

as $K \to \infty$ with zero mean.

Algorithm 1 Fisher algorithm

- 1: Find the initial value $\hat{\beta}_{(0)}$ using generalized linear model **GLM ()**
- 2: Estimate $\hat{\alpha}$ via Pearson residual (See Table 1.2)
- 3: Estimate \hat{G} via Nonlinear Mixed-Effects Models method **nlme** ()
- 4: For given $\hat{\alpha}$ and \hat{G} find the matrix

$$\tilde{V}_{i_{\text{GEEM}}} = V_{i_{\text{GEEM}}}(\beta, \hat{\alpha}, \hat{G})$$

5: Update

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + \left(\sum_{i=1}^{K} \left(D_{i_{\text{GEEM}}}(\hat{\beta}^{(t)}, \hat{\alpha}, \hat{G}) \right)' \left(\tilde{V}_{i_{\text{GEEM}}}(\hat{\beta}^{(t)}, \hat{\alpha}, \hat{G}) \right)^{-1} \left(D_{i_{\text{GEEM}}}(\hat{\beta}^{(t)}, \hat{\alpha}, \hat{G}) \right) \right)^{-1} \\ \times \left(\sum_{i=1}^{K} \left(D_{i_{\text{GEEM}}}(\hat{\beta}^{(t)}, \hat{\alpha}, \hat{G}) \right)' \left(\tilde{V}_{i_{\text{GEEM}}}(\hat{\beta}^{(t)}, \hat{\alpha}, \hat{G}) \right)^{-1} \left(S_{i_{\text{GEEM}}}(\hat{\beta}^{(t)}, \hat{\alpha}, \hat{G}) \right) \right)^{-1}$$

- 6: Evaluate convergence using changes $||\hat{\beta}^{(t+1)} \hat{\beta}^{(t)}||$
- 7: Repeat steps (2) (6) until criterion is satisfied

Proof. The proof is similar to Liang and Zeger (1986). Let $\alpha^* = \hat{\alpha}(\beta, \hat{G})$ and

$$E(K^{-1}\sum_{i=1}^{K} \bigtriangledown^{2} U_{i_{\text{GEEM}}}(\beta, \alpha^{*})) = K^{-1}\sum_{i=1}^{K} \frac{\partial \mu_{i}}{\partial \beta} V_{i_{\text{GEEM}}}^{-1} \frac{\partial \mu_{i}}{\partial \beta}$$

$$= K^{-1}\sum_{i=1}^{K} D_{i_{\text{GEEM}}}^{'} V_{i_{\text{GEEM}}}^{-1} D_{i_{\text{GEEM}}}$$
(2.2.10)

be the expected Hessian matrix and the information matrix of $K^{-1/2} \sum_{i=1}^{K} U_{i_{\text{GEEM}}}(\beta, \alpha^*)$ is given by the CLT as

$$\lim_{K \to \infty} (K^{-1} \sum_{i=1}^{K} D'_{i_{\text{GEEM}}} V_{i_{\text{GEEM}}}^{-1} \operatorname{Cov}(Y_i) V_{i_{\text{GEEM}}}^{-1} D'_{i_{\text{GEEM}}})$$
(2.2.11)

Gourieroux, Monfort, and Trognon (1984). Under regularity conditions, it can be shown by Taylor series that $K^{-1/2}(\hat{\beta} - \beta)$ can be approximated by

$$\left(K^{-1}\sum_{i=1}^{K} - \nabla^2 U_{i_{\text{GEEM}}}(\beta, \alpha^*)\right)^{-1} \left(K^{-1/2}\sum_{i=1}^{K} U_{i_{\text{GEEM}}}(\beta, \alpha^*)\right)$$
(2.2.12)
where the first term in 3.1.17

$$\nabla^2 U_{i_{\text{GEEM}}}(\beta, \alpha^*) = \frac{\partial}{\partial \beta} U_{i_{\text{GEEM}}}(\beta, \alpha^*) + \frac{\partial}{\partial \alpha^*} U_{i_{\text{GEEM}}}(\beta, \alpha^*) \frac{\partial}{\partial \beta} \alpha^*(\beta)$$
(2.2.13)

and the second term in 3.1.17

$$K^{-1/2} \sum_{i=1}^{K} U_{i_{\text{GEEM}}}(\beta, \alpha^{*}) = K^{-1/2} \sum U_{i_{\text{GEEM}}}(\beta, \alpha) + K^{-1} \sum_{i=1}^{K} \frac{\partial}{\partial \alpha} U_{i_{\text{GEEM}}}(\beta, \alpha) K^{-1/2}(\alpha^{*} - \alpha) + o_{p}(1)$$
(2.2.14)

The second term in 3.1.18 is free of Y_i and therefore $\frac{\partial}{\partial \alpha^*} U_{i_{\text{GEEM}}}(\beta, \alpha^*)$ is $o_p(1)$ and $\frac{\partial}{\partial \beta} \alpha^*(\beta)$ is $o_p(1)$. Then, the remaining two terms by LLN have equivalent asymptotic distribution with zero mean and co-variance matrix as in 3.1.16. Similarly, for the second term of 3.1.19 and the remaining two terms by CLT converge to the same limit which is the expected Hessian matrix 3.1.15 and this completes the desired result.

2.3 GEEM for count longitudinal data

Let $Y_i = (y_{i1}, \dots, y_{in_i})'$ be the $n_i \times 1$ response vector for the *i*-th subject and $X_i = (x_{i1}, \dots, x_{in_i})'$ be the $n_i \times p$ matrix of covariate values for the *i*-th subject $i = 1, \dots, K$ corresponding to the fixed effects $\beta \in R^p$, Z_i is the $n_i \times q$ design matrix corresponding to the random effects $b_i \in R^q$, Y_i is generated from Poisson distribution

$$Y_i \sim poisson(\lambda_i) \tag{2.3.1}$$

with the conditional mean and variance

$$g(\lambda_i) = g(E(Y_i|b_i) = X'_i\beta + Z'_ib$$
(2.3.2)

where g = log is the link function and

$$b_i \sim N(0, G) \tag{2.3.3}$$

2.3.1 Marginal means, variances and covariances

The marginal mean $E(y_{ij}) = \exp(x'_{ij}\beta + \frac{1}{2}x'_{ij}Gx_{ij}) = \mu_i$ and variance for Poisson distribution have been derived by Molenberghs et al. (2007) with the assumption that the correlation matrix is identity. The vector of the marginal mean, μ_i depends on explanatory variables X_i , the design matrix Z_i and the variance matrix of the random effects G, through the inverse of the link function.

The working covariance matrix given by

$$\operatorname{Cov}(Y_i) \approx \operatorname{Cov}[\operatorname{E}(Y_i|b_i)] + \operatorname{E}[\operatorname{Cov}(Y_i|b_i)]$$

$$= M_i \left(\exp(Z'_i G Z_i) - J_{ni} \right) M_i + M_i^{1/2} R_i M_i^{1/2}$$
(2.3.4)

OR,

$$V_{i_{GEEM}} = \text{Cov}[\text{E}(Y_i|b_i)] + \text{E}[\text{Cov}(Y_i|b_i)]$$

$$= M_i \left(\exp(Z'_i GZ_i) - J_{ni} \right) M_i + M_i^{1/2} R_i M_i^{1/2}$$
(2.3.5)

where $M_i = diag\{\mu_i\}$, J_{ni} is $n_i \times n_i$ matrix with all elements are ones and R_i is the working correlation matrix. The working correlation matrix R_i should be chosen carefully. The generalized estimating equations for the parameters $\theta = (\beta', b')'$ is given by 2.2.4 with $V_{i_{GEEM}}$ as defined in 2.3.5.

2.4 Simulation study

After proposing the first-order generalized estimating equation for mixed models, simulation studies are needed to investigate the finite sample performance of the proposed method in terms of bias and efficiency. Then, compare the simulated outcomes with those for existing elected methods. The language of **R** have been used for all the generation and the calculation of the simulation data in this dissertation. The **R** codes used for this purpose are available in the appendix. The

references of **R** codes are Xu (2013) and Pavlou (2012).

We are interested in the performance of simulating correlated longitudinal count data responses with known correlation structure. Particularly, under the following cases: first, the performance of the newly estimating procedure based on different design structures of the working correlation matrix; second, the comparison of the simulated outcomes with those for some existing methods. We consider two scenarios: (i) high longitudinal correlated count responses with correlation parameter α close to 1; (ii) medium longitudinal correlated count responses with correlation parameter α ranges between 0.4 to 0.6.

For each scenario we generate correlated Poisson random variables for K correlated subjects. That is, we generate a dataset from the underlying model 2.4.1. Combining the fixed effects and random effects gives

$$\eta = X_i'\beta + Z_i'b_i$$

which form a linear predictor (See section 1.2.1). The underlying model is

$$y = \eta + \epsilon = X'_i \beta + Z'_i b_i + \epsilon \tag{2.4.1}$$

where ϵ is disturbance term. We choose the following parameters: the coefficients parameter β ; the correlation parameter α ; the variance of the random effects G corresponding to high and medium correlated simulated dataset.

2.4.1 Simulation 1

The simulation study is needed to investigate the finite sample performance of the proposed method in term of bias and efficiency.

Scenario 1

For this scenario, the simulated dataset is highly correlated. We consider K = 50, 100, 150, and set $n_i = 4$ under the cases below. In each case consider the true model with p = 4 for fixed effects parameters with true parameter vector $\beta = (3, .5, 1, .2)$, and q = 2 for random effects. case (I) First, using the packages corcounts and mmm generate the responses $Y_i \sim \text{Poisson}(\lambda_i)$ for the *i*-th cluster with the correlation matrix

$$R_i(\alpha) = \begin{bmatrix} 1.000 & 0.850 & 0.850 & 0.850 \\ .0850 & 1.000 & 0.850 & 0.850 \\ 0.850 & 0.850 & 1.000 & 0.850 \\ 0.850 & 0.850 & 0.850 & 1.000 \end{bmatrix}$$
(2.4.2)

where the structure of the correlation matrix is exchangeable (compound symmetry). Then, generate the covariates matrix X_i as

$$x_{ij} = \begin{cases} 0 & \text{if } j = 0 \\ 1 & \text{if } j = 1 \\ 2 & \text{if } j = 2 \\ 3 & \text{if } j = 3 \end{cases}$$

$$x_i = \begin{cases} 0 & \text{for the first } K/2 \text{ clusters} \\ 1 & \text{for the remaining clusters} \end{cases}$$

$$(2.4.4)$$

Assume that, $z_{ij} = x_{ij}$. When $x_i = 1$ the matrices X_i and Z_i are as following

$$X_{i}^{'} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 2 \\ 1 & 3 & 1 & 3 \end{bmatrix} \qquad Z_{i}^{'} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$$

which produces the model

$$E(y_{ij}|b_i) = \exp\begin{cases} \beta_0 + b_{0i} + \beta_2 & \text{if } j = 0\\ \beta_0 + b_{0i} + \beta_1 + b_{1i} + \beta_2 + \beta_3 & \text{if } j = 1, 2, 3. \end{cases}$$
(2.4.5)

The general form of the model when $x_i = 1$ can be written as

$$\log(\mathbf{E}(Y_i|b_i)) = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 2 \\ 1 & 3 & 1 & 3 \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \times \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix}$$

When $x_i = 0$ the matrices X_i and Z_i are as following

$$X_{i}^{\prime} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 3 & 0 & 0 \end{bmatrix} \qquad Z_{i}^{\prime} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$$

which produces the model

$$\mathbf{E}(y_{ij}|b_i) = \exp\begin{cases} \beta_0 + b_{0i} + \beta_2 & \text{if } j = 0\\ \beta_0 + b_{0i} + \beta_1 + b_{1i} + \beta_2 + \beta_3 & \text{if } j = 1, 2, 3. \end{cases}$$
(2.4.6)

The general form of the model when $x_i = 0$ can be written as

$$\log(\mathbf{E}(Y_i|b_i)) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 3 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \times \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix}$$
$$\begin{pmatrix} \beta_0 + b_{0i} & \text{if } j = 0 \end{bmatrix}$$

$$E(y_{ij}|b_i) = \exp \begin{cases} \beta_0 + b_{0i} & \text{if } j = 0\\ \beta_0 + b_{0i} + \beta_1 + b_{1i} & \text{if } j = 1, 2, 3. \end{cases}$$
(2.4.7)

Hence, the model for this simulated data is given by

$$\log(\mathsf{E}(y_{ij}|b_i)) = \beta_0 + \beta_1 x_{ij} + \beta_2 x_i + \beta_3 x_{ij} * x_i + b_{0i} + b_{1i} z_{ij}$$
(2.4.8)

with $i = 1, \dots, K$ and j = 0, 1, 2, 3.

case (II) The simulation study in the previous case studied the behavior of the finite samples K = 50, 100, 150 with the highly correlated data that were generated under compound symmetry structure of the correlation matrix $R_i(\alpha)$. In this case, we repeat the same process mentioned above but switching the correlation structure to autoregressive order 1 (AR(1)). Generate the correlated data following the model in 2.4.1 with true correlation matrix when

$$\alpha = 0.929 \tag{2.4.9}$$

case (III) Repeat the exact processes with generating highly correlated data with the unstructured cor-

relation matrix

$$R_{i}(\alpha) = \begin{bmatrix} 1.000 & 0.900 & 0.800 & 0.700 \\ 0.900 & 1.00 & 0.830 & 0.850 \\ 0.800 & 0.830 & 1.000 & 0.96 \\ 0.700 & 0.850 & 0.960 & 1.000 \end{bmatrix}$$
(2.4.10)

Scenario 2

In this scenario the longitudinal data is not highly correlated but we are still assuming the observations for each cluster are correlated. That is, we are assuming the correlation parameter α is between 0.3 and 0.5. Consider K = 50, 100, 150, and we set $n_i = 4$ under the cases below. In each case consider the true model with p = 4 for fixed effects parameters with true parameter vector $\beta = (3.00, 0.50, 1.00, .2)$, and q = 2 for random effects.

case (I) Generate the correlated data following the model in 2.4.1 with true exchangeable (compound symmetry) correlation matrix

$$R_i(\alpha) = \begin{bmatrix} 1.000 & 0.459 & 0.459 & 0.459 \\ 0.459 & 1.000 & 0.500 & 0.459 \\ 0.459 & 0.459 & 1.000 & 0.459 \\ 0.459 & 0.459 & 0.459 & 1.000 \end{bmatrix}$$
(2.4.11)

case (II) Generate the correlated data following the model in 2.4.1 with true autoregressive order 1 (AR(1)) correlation matrix when

$$\alpha = 0.300$$
 (2.4.12)

case (III) Repeat the exact processes by generating correlated data with the unstructured correlation

matrix given by

$$R_i(\alpha) = \begin{bmatrix} 1.000 & 0.400 & 0.600 & 0.700 \\ 0.400 & 1.000 & 0.600 & 0.370 \\ 0.600 & 0.600 & 1.000 & 0.600 \\ 0.700 & 0.370 & 0.600 & 1.000 \end{bmatrix}$$
(2.4.13)

2.4.2 Simulation 2

We compare the simulated outcomes with those of existing selection methods such as GEE, GLM and GLMM. We compare our proposed method to GEE since it is the basis of our method. GEE proposed to handle only marginal model and our model is conditional; hence we also compare our method with some selected approaches that are applicable to conditional models.

2.4.3 Plots for simulated data when K = 50

Visualizing the data before finding and analyzing the estimates is very helpful. Figures (2.1), (2.5) and (2.9) are box-plots of the observations while figures (2.2), (2.6) and (2.10) are the boxplots of the log of the observations for K = 50, K = 100 and K = 150, respectively. Figures (2.3) and (2.8) show the trajectory of each cluster individually figures (2.4), (2.7) and (2.11) show the trajectory of all the clusters in one profile.



Figure 2.1: Boxplots of numbers of observations when K = 50.



Figure 2.2: Boxplots of log of numbers of observations for K = 50.









2.4.4 Plots for simulated data when K = 100



Figure 2.5: Boxplots of numbers of observations when K = 100.



Figure 2.6: Boxplots of log of numbers of observations for K = 100.





10	20	30	40	50	60	70	80	06	100	0 1 2 3
σ	19	29	30	49	20	69	79	80	66	0 1 2 3
α		28	38	48	28	88	78	88	6	0 1 2 3
~	17	27	37	47	57	67	77	87	97	0 1 2 3
00	16	26	36	46	56	66	76	86	- 30 -	1 2 3
<u>م</u>	10 	25	35	45	55	65	75	85	95	0 1 2 3 ×
4	14	24	34	44	54	64	74	84	94	0 1 2 3
с С	13	23	33	43	53	63	73	83	33	0 1 2 3
2	12	22	32	42	52	62	72	82	92	0 1 2 3
- (=	21	31	41	21	61	4	81	6	0 1 2 3
404-	4.04	404	404	404- (0000)	Γ\ (104- (104- (104- (104-))	40 4 (0000)	404	404-	404- 00000	



2.4.5 Plots for simulated data when K = 150



Figure 2.9: Boxplots of numbers of observations when K = 150.



Figure 2.10: Boxplots of log of numbers of observations for K = 150.





13	26	^ 30	52		65	78	\langle	91	\langle	104	\langle	117	\langle	130	\langle	143		0 - -	
12	25	38	51		64	27		06	\langle	103	\langle	116	\langle	129	\langle	142	- c - c	0 1 1	
5	24	37	50		63	76		89	\langle	102	\langle	115	\langle	128	$\left\langle \right\rangle$	141		0 - 0	
10	23	36	49		62	75		88	\langle	101		114	\langle	127	\langle	140			
σ	22	35	48		61	74		87	\langle	100	\langle	113	\langle	126	\langle	139			
∞	21	34	47		60	73		86	\langle	66	\langle	112	\langle	125	\langle	138		0 7 1	
~	20	33	46		1	72		85	\langle	98	\langle	111	\langle	124	\langle	137	$\left\langle \right\rangle$	150	×
0	6	32	45		58	71		84	\langle	97	\langle	110	\langle	123	\langle	136	$\left\langle \right\rangle$	149	0-1- 0
2		31	44		57	70		83	\langle	96	\langle	109	\langle	122	\langle	135		148	- 2- - 2-
4	17	30	43		56	69		82	\langle	95	\langle	108	\langle	121	\langle	134	\langle	147	0-1- 0
m	16	29	42		55	68		81	\langle	94		107	\langle	120	\langle	133	\langle	146	0-1- 0
7	15	28	41		54	67		80		93	\langle	106	\langle	119	\langle	132	\langle	145	0
-	4	27	40		23	66		79	\langle	92	\langle	105	\langle	118	\langle	131		144	0 - 7 - 0
788		266		368 198	280 280	00 - -	388 788	۱۱		0				000					77- 1990 1990



Table 2.1: Parameter estimates (standard deviations in parentheses) of the true beta values (3.00, 0.50, 1.00, 0.20) when number of clusters K = 50, 100 and 150 with $n_i = 4$ per each cluster, is compared for: the newly proposed method (GEEM), GEE Liang and Zeger (1986), GLM McCullagh (1984) and GLMM method, for 5,000 simulations applying the model 2.4.1 with correlation matrix as shown in 2.4.13

.

K	Parameters	GEEM	GEE	GLM	GLMM
	eta_0	4.360(.99841)	4.361(.99980)	4.371(.99980)	4.093(1.09235)
K=50	β_1	0.654(.00529)	0.651(.00587)	0.647(.00587)	0.647(.00861)
	β_2	0.664(.02157)	0.671(.02233)	0.667(.02233)	0.529(.04277)
	eta_3	0.222(.00540)	0.223(.00518)	0.224(.00518)	0.244(.00850)
	eta_0	4.225(.99930)	4.223(.99868)	4.229(.99885)	4.002(1.87739)
K=100	β_1	0.691(.00763)	0.691(.00781)	0.690(.00765)	0.690(.00956)
	β_2	0.912(.02050)	0.911(.02000)	0.909(.01998)	1.736(0.0731)
	eta_3	0.187(.00893)	0.190(.00892)	0.190(.00898)	0.198(.01201)
	eta_0	4.242(.97899)	4.241(.98066)	4.242(.98066)	4.042(.98568)
K=150	β_1	0.686(.00367)	0.686(.00335)	0.686(.00335)	0.685(.00366)
	β_2	0.925(.01733)	0.924(.01788)	0.924(.01788)	0.861(.01918)
	eta_3	0.197(.00459)	0.198(.00443)	0.198(.00443)	0.198(00464)

Table 2.2: Parameter estimates (standard deviations in parentheses) of the true beta values (3.00, 0.50, 1.00, 0.20) when number of clusters K = 50, 100 and 150 with $n_i = 4$ per each cluster, is compared for: the newly proposed method (GEEM), GEE Liang and Zeger (1986), GLM McCullagh (1984) and GLMM method, for 5,000 simulations applying the model 2.4.1 with correlation matrix as shown in 2.4.10

•

K	Parameters	GEEM	GEE	GLM	GLMM
	β_0	4.392(.97769)	4.392(.97855)	4.401(.97730)	4.003(1.02969)
K=50	β_1	0.654(.00775)	0.650(.00828)	0.648(.00860)	0.688(.01861)
	β_2	0.665(.02776)	0.667(.02684)	0.664(.02763)	0.520(0.09627)
	eta_3	0.213(.00816)	0.214(.00829)	0.214(.00851)	0.214(.02500)
	eta_0	4.254(.99784)	4.252(.99897)	4.253(.99994)	3.992(1.87739)
K=100	β_1	0.685(.00583)	0.685(.00578)	0.685(.00566)	0.680(.02600)
	β_2	0.898(.01843)	0.899(.01843)	0.899(.01739)	1.536(.07198)
	eta_3	0.192(.00560)	0.191(.00568)	0.190(.00595)	0.198(.01120)
	eta_0	4.269(.97485)	4.268(.97550)	4.269(.97553)	4.062(1.00517)
K=150	β_1	0.681(.00745)	0.681(.00746)	0.681(.00747)	0.681(.00946)
	β_2	0.918(.02080)	0.918(.02028)	0.919(.02134)	0.862(.03991)
	eta_3	0.198(.00675)	0.197(.00667)	0.197(.00686)	0.197(.00886)

Table 2.3: Parameter estimates (standard deviations in parentheses) of the true beta values (3.00, 0.50, 1.00, 0.20) when number of clusters K = 50, 100 and 150 with $n_i = 4$ per each cluster, is compared for: the newly proposed method (GEEM), GEE Liang and Zeger (1986), GLM McCullagh (1984) and GLMM method, for 5,000 simulations applying the model 2.4.1 with correlation matrix as shown in 2.4.9

•

K	Parameters	GEEM	GEE	GLM	GLMM
	β_0	4.407(.97396)	4.407(.97362)	4.411(.97308)	4.008(.98105)
K=50	β_1	0.647(.00752)	0.645(.00806)	0.644(.00834)	0.644(.01340)
	β_2	0.655(.02290)	0.654(.02332)	0.653(.02503)	0.504(.02712)
	eta_3	0.217(.00694)	0.218(.00726)	0.219(.00754)	0.219(.00954)
	eta_0	4.265(1.00192)	4.265(1.00185)	4.264(1.00299)	4.040(1.02149)
K=100	β_1	0.682(.00658)	0.681(.00664)	0.681(.00669)	0.679(.02706)
	β_2	0.889(.01677)	0.888(.01532)	0.887(.01486)	1.333(.09209)
	eta_3	0.194(.00557)	0.194(.00564)	0.195(.00581)	0.198(0.00804)
	eta_0	4.259(.97736)	4.259(.97596)	4.261(.97419)	4.100(.98212)
K=150	β_1	0.683(.00474)	0.682(.00537)	0.682(.00570)	0.688(.00700)
	β_2	0.921(.01794)	0.922(.01813)	0.921(.01962)	0.858(.02354)
	eta_3	0.197(.00442)	0.197(.00467)	0.197(.00494)	0.197(.00994)

Table 2.4: Parameter estimates (standard deviations in parentheses) of the true beta values (3.00, 0.50, 1.00, 0.20) when number of clusters K = 50, 100 and 150 with $n_i = 4$ per each cluster, is compared for: the newly proposed method (GEEM), GEE Liang and Zeger (1986), GLM McCullagh (1984) and GLMM method, for 5,000 simulations applying the model 2.4.1 with correlation matrix as shown in 2.4.11

•

K	Parameters	GEEM	GEE	GLM	GLMM
	β_0	4.352(.98275)	4.352(.98876)	4.368(.98726)	4.007(1.67351)
K=50	β_1	0.631(.00925)	0.624(.00974)	0.621(.00965)	0.062(.00950)
	β_2	0.674(.01779)	0.674(.01976)	0.670(.01813)	0.558(.01816)
	eta_3	0.241(.00866)	0.245(.00884)	0.246(.00868)	0.246(.0088)
	eta_0	4.223(.99228)	4.352(.98876)	4.233(.99329)	4.311(1.07446)
K=100	β_1	0.682(.00573)	0.624(.00774)	0.678(.00627)	0.678(.00770)
	β_2	0.909(.01922)	0.974(.01976)	0.904(.01924)	1.633(.04749)
	eta_3	0.197(.00702)	0.245(.00784)	0.200(.00784)	0.211(0.00788)
	eta_0	4.252(.97341)	4.250(.97706)	4.256(.97797)	4.252(1.03558)
K=150	β_1	0.678(.00670)	0.677(.00687)	0.676(.00686)	0.676(.00586)
	β_2	0.920(.01824)	0.921(.01827)	0.920(.01896)	0.858(.03260)
	eta_3	0.202(.00564)	0.203(.00581)	0.203(.00586)	0.203(.00596)

As mentioned above, we generate the data based on the model 2.4.1 with various choices of correlation structures ranging from highly correlated data to medium correlated data. In this chapter our interest focuses on the regression coefficients and estimate the fixed effects β using the estimating equations 2.2.6.

Generalized estimating equations method Liang and Zeger (1986) is known to provide consistent estimates. We estimate the fixed effects β using the estimating equations 2.2.6 and compare the estimates to those from generalized estimating equations method using the function **gee()**. In addition, we compare the estimates on our newly propsed model to the estimates of some selected methods such as GLM, GLMM using the functions **glm()**, and **glmmML()**, receptively.

Tables 2.1 and 2.2 show the results of the first scenario. Tables 2.3 and 2.4 show the results of the second scenario. In general, Tables 2.1-2.4 show that the generalized estimating equations for mixed model GEEM 2.2.6 perform very well under both scenarios.

GLM and GEE methods are known to provide content estimates. In comparison to these two methods it can be seen that all the estimates of fixed effects parameter $\hat{\beta}$ of the all methods analyzed are very close to those provided by generalized estimating equations method for all choices of sample sizes (i.e. for K = 50, K = 100 and K = 150). In addition, the estimated provided by our newly propsed method are slightly closer to the true beta values in almost all ceases. Although there are slight differences between the standard deviations provided by the methods shown in Tables 2.1-2.4 but they are very close. It can be seen that among the four methods GLMM produces estimators with higher standard deviations. While our proposed method produces the least standard deviations among all methods in almost all cases. We conclude that GEEM is the most efficient between content estimators, here.

CHAPTER 3 SECOND-ORDER GENERALIZED ESTIMATING EQUATIONS FOR MIXED MODELS

3.1 GEE2 for mixed models

GEE2 approach is another class of estimating equations proposed by Prentice (1988) for longitudinal binary data and devolved by Prentice and Zhao (1991) for longitudinal data for any distribution in exponential family. This system has been used to estimate the parameter β and the correlation estimates α for marginal models simultaneously via a joint estimating equations Ziegler, Kastner, Grömping, and Blettner (1996).

For some longitudinal studies the correlation structure is of scientific interest, for instance, epidemiological studies. The correlation structure has been studied by numerous authors as Prentice and Zhao (1991), Carey, Zeger, and Diggle (1993), Yi and Cook (2002) among others but for marginal distributions. We extend their method for conditional distributions in two ways. First way, assuming we are interested in the variability between subjects in addition to the coefficient parameters. Alternatively, assume that we are interested in the coefficient parameters and the association structure.

Assume that the random effects b_i is following normal distribution with zero mean and 2×2 variance matrix G.

$$b_i \text{ i.i.d } N(0,G)$$
 (3.1.1)

Let $Y_i = (y_{i1}, \dots, y_{in_i})'$ be the $n_i \times 1$ response vector and $X_i = (x_{i1}, \dots, x_{in_i})'$ be the $n_i \times p$ matrix of covariate values for the *i*-th subject $i = 1, \dots, K$ corresponding to the fixed effects $\beta \in R^p$, Z_i is the $n_i \times q$ design matrix corresponding to the random effects $b_i \in R^q$, y_{ij} , is generated from a distribution in exponential family with conditional mean

$$g(\mathbf{E}(Y_i|b_i)) = X'_i\beta + Z'_ib_i$$
(3.1.2)

where g is a link function. The random effects b_i are assumed to be mutually independent following the normal distribution with zero mean and covariance matrix G. The exact covariance matrix of Y_i is unknown but if the working correlation matrix $R(\alpha)$ is chosen correctly then the approximation of the covariance matrix of Y_i is

$$V_{i_{\text{GEEM}}} = \text{Cov}[E(Y_i|b_i)] + E[\text{Cov}(Y_i|b_i)]$$

= $Z'_i GZ_i + A_i^{1/2} R_i(\alpha) A_i^{1/2}$ (3.1.3)

In other words,

$$Cov(Y_i) \approx Z'_i G Z_i + A_i^{1/2} R_i(\alpha) A_i^{1/2}$$
 (3.1.4)

where matrix A_i is a diagonal matrix that consists of the second moment of the model and $R_i(\alpha)$ is a correlation matrix (See section 1.2.3), and therefore the model error is not necessarily independent.

The first part of the second-order system of the generalized estimating equations for the conditional model is as shown in 3.1.3 with the $V_{i_{GEEM}}$. Since, the second-order system consists of a pair of estimating equations we construct an analogue estimating equations to 3.1.3 to solve for two parameters – either (β, b) or (β, α) – while considering the third parameters as a nuisance.

3.1.1 Estimation β and G, simultaneously

Sutradhar and Jowaheer (2003) have used Prentice and Zhao (1991) method to estimate the variance of the random effects. They used it for univariate random effect. We extend it to multi-variate normal random effect. To accomplish this estimation by means of GEE, define

$$T_{i} = (y_{i} - \mu_{i})(y_{i} - \mu_{i})'$$

$$= ((y_{i1} - \mu_{i1})^{2}, (y_{i2} - \mu_{i2})^{2}, \cdots, (y_{ij} - \mu_{ij})^{2},$$

$$(y_{i1} - \mu_{i1})(y_{i2} - \mu_{i2}), \cdots, (y_{i,j-1} - \mu_{i,j-1})(y_{ij} - \mu_{ij}))'$$

$$= (t_{i11}, t_{i22}, \cdots, t_{ijj}, t_{i12}, t_{i23}, \cdots, t_{i,j-1,j})'$$
(3.1.5)

be an $n_i(n_i - 1)/2 + n_i \times 1$ vector with $E(T_i) = \zeta_i$. Then construct a parallel system of estimating equations to 2.2.6 as

$$U_{\text{GEEM2}}[\beta, \alpha(\beta), G] = \sum_{i=1}^{K} \left(\partial \zeta(\beta, \alpha, G) / \partial G \right)' \left(V_{i_{\text{GEEM2}}}(\beta, \alpha, G) \right)^{-1} \left(T_i - \zeta(\beta, \alpha, G) \right)$$
$$= \sum_{i=1}^{K} \left(D_{i_{\text{GEEM2}}}(\beta, \alpha, G) \right)' \left(V_{i_{\text{GEEM2}}}(\beta, \alpha, G) \right)^{-1} \left(S_{i_{\text{GEEM2}}}(\beta, \alpha, G) \right) = 0$$
(3.1.6)

where $S_{i_{\text{GEEM2}}} = T_i - \zeta_i$, $D_{i_{\text{GEEM2}}} = \partial \zeta_i / \partial G$, and $V_{i_{\text{GEEM2}}}$ is the working covariance matrix of the vector T_i and might contain information about higher moments (i.e third and fourth moments) of y_i Lo, Fung, and Zhu (2007).

For constructing the $n_i(n_i - 1)/2 + n_i \times n_i(n_i - 1)/2 + n_i$ matrix $V_{i_{\text{GEEM2}}}$ some authors used the law of total probability covariance as

$$Cov(t_{ij}^{2}, t_{il}t_{im}) = Cov[E(t_{ij}^{2}|b_{i})] + E[Cov(t_{il}t_{im}|b_{i})]$$
(3.1.7)

$$Cov(t_{ij}^2, t_{il}^2) = Cov[E(t_{ij}^2|b_i)] + E[Cov(t_{il}^2|b_i)]$$
(3.1.8)

$$\operatorname{Cov}(t_{ij}t_{ik}, t_{il}t_{im}) = \operatorname{Cov}[\operatorname{E}(t_{ij}t_{ik}|b_i)] + \operatorname{E}[\operatorname{Cov}(t_{il}t_{im}|b_i)]$$
(3.1.9)

Rudary (2009). Since the second order $V_{i_{\text{GEEM2}}}$ matrix consist of higher moments order such as third and fourth moments as 3.1.7-3.1.9 which are highly unstable and to this reason we consider $V_{i_{\text{GEEM2}}}$ is $n_i(n_i - 1)/2 + n_i \times n_i(n_i - 1)/2 + n_i$ identity matrix Wakefield (2009). We have built the parallel estimating equations in 3.1.6. To simultaneously model β and G we use the estimating equation 2.2.6 and its parallel 3.1.6 to form one system given by

$$U_{\text{full}}(\beta, \alpha(\beta), G) = \sum_{i=1}^{K} \begin{bmatrix} D_{i11} & D_{i12} \\ D_{i21} & D_{i22} \end{bmatrix}' \begin{bmatrix} V_{i11} & V_{i12} \\ V_{i21} & V_{i22} \end{bmatrix}^{-1} \begin{bmatrix} S_{i1} \\ S_{i2} \end{bmatrix}$$
$$= \sum_{i=1}^{K} \left(D_{i_{\text{full}}}(\beta, \alpha(\beta), G) \right)' \left(V_{i_{\text{full}}}(\beta, \alpha(\beta), G) \right)^{-1} \left(S_{i_{\text{full}}}(\beta, \alpha(\beta), G) \right) = 0$$
(3.1.10)

where,

$$\begin{split} D_{i_{\text{full}}}(\beta, \alpha(\beta), G) &= \begin{bmatrix} D_{i11} & D_{i12} \\ D_{i21} & D_{i22} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mu_i}{\partial \beta} & \frac{\partial \mu_i}{\partial G} \\ \frac{\partial \zeta_i}{\partial \beta} & \frac{\partial \zeta_i}{\partial G} \end{bmatrix} \\ V_{i_{\text{full}}}(\beta, \alpha(\beta), G) &= \begin{bmatrix} V_{i11} & V_{i12} \\ V_{i21} & V_{i22} \end{bmatrix} = \begin{bmatrix} \text{Var}(Y_i) & \text{Cov}(Y_i, T_i) \\ \text{Cov}(T_i, Y_i) & \text{Var}(T_i) \end{bmatrix} \\ S_{i_{\text{full}}}(\beta, \alpha(\beta), G) &= \begin{bmatrix} S_{i1} \\ S_{i2} \end{bmatrix} = \begin{bmatrix} Y_i - \mu_i \\ T_i - \zeta_i \end{bmatrix} \end{split}$$

The vector $S_{i_{\text{full}}}(\beta, \alpha(\beta), G)$ consist of the $S_{i1} = Y_i - \mu_i$ and $S_{i2} = T_i - \zeta_i$. The form of GEEM2 that we have proposed in 3.1.10 is the general form. Prentice and Zhao (1991) provided three special structures to 3.1.10, independence structure, normal structure and normal structure with common third and fourth order correlation (See section 1.3). Our model is complicated enough and have two layers of correlations within- and between-subject correlations. The general structure in 3.1.10 is not plausible due to the complication of interpretation. Therefore, choosing independence

structure reduces the GEEM2 3.1.10 to the following reduced GEEM2 form

$$U_{\text{redu}}(\beta, \alpha(\beta), G) = \sum_{i=1}^{K} \begin{bmatrix} D_{i11} & 0 \\ 0 & D_{i22} \end{bmatrix}' \begin{bmatrix} V_{i11} & 0 \\ 0 & V_{i22} \end{bmatrix}^{-1} \begin{bmatrix} S_{i1} \\ S_{i2} \end{bmatrix}$$
$$= \sum_{i=1}^{K} \left(D_{i_{\text{redu}(G)}}(\beta, \alpha(\beta), G) \right)' \left(V_{i_{\text{redu}(G)}}(\beta, \alpha(\beta), G) \right)^{-1} \left(S_{i_{\text{redu}(G)}}(\beta, \alpha(\beta), G) \right) = 0$$
(3.1.11)

where,

$$\begin{split} D_{i_{\text{redu}(G)}} &= \begin{bmatrix} D_{i11} & 0\\ 0 & D_{i22} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mu_i}{\partial \beta} & 0\\ 0 & \frac{\partial \zeta_i}{\partial G} \end{bmatrix} \\ V_{i_{\text{redu}(G)}} &= \begin{bmatrix} V_{i11} & 0\\ 0 & V_{i22} \end{bmatrix} = \begin{bmatrix} \text{Var}(Y_i) & 0\\ 0 & \text{Var}(T_i) \end{bmatrix} \\ S_{i_{\text{redu}(G)}} &= \begin{bmatrix} S_{i1}\\ S_{i2} \end{bmatrix} = \begin{bmatrix} Y_i - \mu_i\\ T_i - \zeta_i \end{bmatrix} \end{split}$$

The matrices D_{i11} , D_{i22} , V_{i11} and D_{i22} remain as defined above in the general form of GEE2 but because we are assuming independence structure the matrices D_{i12} , D_{i21} , V_{i12} and D_{i21} are all zeros.

- 1: Find initial values of $\hat{\beta}$ using **GLM()**
- 2: Find initial values of the variance matrix G through the function **nlme()**.
- 3: Estimate $\hat{\alpha}$ via Pearson residual (See Section 1.2)
- 4: Iterate until convergence

$$\begin{bmatrix} \hat{\beta}^{(t+1)} \\ \hat{G}^{(t+1)} \end{bmatrix} = \begin{bmatrix} \beta^{(t)} \\ G^{(t)} \end{bmatrix} - \left[\sum_{i=1}^{K} (D_{i_{\text{redu}(G)}}(\beta^{(t)}, \hat{\alpha}, G^{(t)}))' (V_{i_{\text{redu}(G)}}(\beta^{(t)}, \hat{\alpha}, G^{(t)}))^{-1} D_{i_{\text{redu}(G)}}(\beta^{(t)}, \hat{\alpha}, G^{(t)}) \right]^{-1} \\ \times \left[\sum_{i=1}^{K} (D_{i_{\text{redu}(G)}}(\beta^{(t)}, \hat{\alpha}, G^{(t)}))' (V_{i_{\text{redu}(G)}}(\beta^{(t)}, \hat{\alpha}, G^{(t)}))^{-1} S_{i_{\text{redu}(G)}}(\beta^{(t)}, \hat{\alpha}, G^{(t)}) \right]^{-1}$$

- 5: Evaluate convergence using $||\hat{\beta}^{(t+1)} \hat{\beta}^{(t)}||$
- 6: Repeat steps (2) (5) until criterion is satisfied

Theorem 3.1.12. The estimators $(\hat{\beta}, \hat{G})$ of (β, G) are consistent and is asymptotically normal

$$\sqrt{K}\left((\hat{\beta}-\beta),(\hat{G}-G)\right)' \to N(0,H)$$
 (3.1.13)

where the asymptotic covariance matrix H is

$$H = \lim_{K \to \infty} K \left(\sum_{i=1}^{K} D'_{i_{full}} V_{i_{full}}^{-1} D_{i_{full}} \right)^{-1} \left(\sum_{i=1}^{K} D'_{i_{full}} V_{i_{full}}^{-1} Cov(Y_i) V_{i_{full}}^{-1} D_{i_{full}} \right) \left(\sum_{i=1}^{K} D'_{i_{full}} V_{i_{full}}^{-1} D_{i_{full}} \right)^{-1}$$
(3.1.14)

as $K \to \infty$ with zero mean.

Proof. The proof follows the same lines as in proof of Theorem 2.2.8. and

$$\mathbf{E}(K^{-1}\sum_{i=1}^{K} \bigtriangledown^{2} U_{i_{\text{GEEM2}}}(\beta, \alpha^{*})) = K^{-1}\sum_{i=1}^{K} \frac{\partial \mu_{i}}{\partial \beta} V_{i_{\text{GEEM2}}}^{-1} \frac{\partial \mu_{i}}{\partial \beta}$$

$$= K^{-1}\sum_{i=1}^{K} D_{i_{\text{GEEM2}}}^{'} V_{i_{\text{GEEM2}}}^{-1} D_{i_{\text{GEEM2}}}$$
(3.1.15)

be the expected Hessian matrix and the information matrix of $K^{-1/2} \sum_{i=1}^{K} U_{i_{\text{GEEM2}}}(\beta, \alpha^*)$) is given

by the CLT as

$$\lim_{K \to \infty} (K^{-1} \sum_{i=1}^{K} D'_{i_{\text{GEEM2}}} V_{i_{\text{GEEM2}}}^{-1} \operatorname{Cov}(Y_i) V_{i_{\text{GEEM2}}}^{-1} D'_{i_{\text{GEEM2}}})$$
(3.1.16)

Gourieroux et al. (1984). Under regularity conditions, it can be shown by Taylor series that $K^{-1/2}(\hat{\beta} - \beta)$ can be approximated by

$$\left(K^{-1}\sum_{i=1}^{K} -\nabla^2 U_{i_{\text{GEEM2}}}(\beta, \alpha^*)\right)^{-1} \left(K^{-1/2}\sum_{i=1}^{K} U_{i_{\text{GEEM2}}}(\beta, \alpha^*)\right)$$
(3.1.17)

where the first term in 3.1.17

$$\nabla^2 U_{i_{\text{GEEM2}}}(\beta, \alpha^*) = \frac{\partial}{\partial \beta} U_{i_{\text{GEEM2}}}(\beta, \alpha^*) + \frac{\partial}{\partial \alpha^*} U_{i_{\text{GEEM2}}}(\beta, \alpha^*) \frac{\partial}{\partial \beta} \alpha^*(\beta)$$
(3.1.18)

and the second term in 3.1.17

$$K^{-1/2} \sum_{i=1}^{K} U_{i_{\text{GEEM2}}}(\beta, \alpha^*) = K^{-1/2} \sum U_{i_{\text{GEEM2}}}(\beta, \alpha) + K^{-1} \sum_{i=1}^{K} \frac{\partial}{\partial \alpha} U_{i_{\text{GEEM2}}}(\beta, \alpha) K^{-1/2}(\alpha^* - \alpha) + o_p(1)$$
(3.1.19)

The second term in 3.1.18 is free of Y_i and therefore $\frac{\partial}{\partial \alpha^*} U_{i_{\text{GEEM2}}}(\beta, \alpha^*)$ is $o_p(1)$ and $\frac{\partial}{\partial \beta} \alpha^*(\beta)$ is $o_p(1)$. Then, the remaining two terms by LLN have equivalent asymptotic distribution with zero mean and co-variance matrix as in 3.1.16. Similarly for the second term of 3.1.19 and the remaining two terms by CLT converge to the same limit which is the expected Hessian matrix 3.1.15 and the completes the desired result.

3.1.2 Estimating β and α , simultaneously

The correlation structure of the matrix R_i is of interest in some longitudinal studies. Prentice and Zhao (1991) proposed a second-order of estimating equation for this propose. Liang, Zeger, and Qaqish (1992) used the phrase GEE2 for the second-order generalized estimating equations which estimate α and β simultaneously. Their method has been used for marginal models, we have used it in the previous section to estimate the coefficient parameters and the covariance of the random effects in the conditional models.

We are primarily interested in β but because the correlation structure may also be of interest in some studies, GEE2 method allow one to model pairwise association in addition to the coefficient parameters. We use the reduced form of GEE2 3.1.11 while considering the covariance of the random effects as nuisance.

The second-order of the generalized estimating equation in this case is given by

$$U_{\text{redu}(\alpha)}(\beta, \alpha(\beta), G) = \sum_{i=1}^{K} \begin{bmatrix} D_{i11} & 0 \\ 0 & D_{i22} \end{bmatrix}' \begin{bmatrix} V_{i11} & 0 \\ 0 & V_{i22} \end{bmatrix}^{-1} \begin{bmatrix} S_{i1} \\ S_{i2} \end{bmatrix}$$
$$= \sum_{i=1}^{K} \left(D_{i_{\text{redu}(\alpha)}}(\beta, \alpha(\beta), G) \right)' \left(V_{i_{\text{redu}(\alpha)}}(\beta, \alpha(\beta), G) \right)^{-1} \left(S_{i_{\text{redu}(\alpha)}}(\beta, \alpha(\beta), G) \right) = 0$$
(3.1.20)

where,

$$D_{i_{\text{redu}(\alpha)}} = \begin{bmatrix} D_{i11} & 0 \\ 0 & D_{i22} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mu_i}{\partial \beta} & 0 \\ 0 & \frac{\partial \zeta_i}{\partial \alpha} \end{bmatrix}$$
$$V_{i_{\text{redu}(\alpha)}} = \begin{bmatrix} V_{i11} & 0 \\ 0 & V_{i22} \end{bmatrix} = \begin{bmatrix} \text{Var}(Y_i) & 0 \\ 0 & \text{Var}(T_i) \end{bmatrix}$$
$$S_{i_{\text{redu}(\alpha)}} = \begin{bmatrix} S_{i1} \\ S_{i2} \end{bmatrix} = \begin{bmatrix} Y_i - \mu_i \\ T_i - \zeta_i \end{bmatrix}$$

The matrices D_{i11} , D_{i22} , V_{i11} and D_{i22} remain as defined above in the general form of GEE2 but because we are assuming independence structure the matrices D_{i12} , D_{i21} , V_{i12} and D_{i21} are all zeros.

- 1: Find initial values of $\hat{\beta}$ using **GLM()**
- 2: Find initial values of the variance matrix α through the function gee () .
- 3: Iterate until convergence

$$\begin{bmatrix} \hat{\beta}^{(t+1)} \\ \hat{\alpha}^{(t+1)} \end{bmatrix} = \begin{bmatrix} \beta^{(t)} \\ \alpha^{(t)} \end{bmatrix} - \left[\sum_{i=1}^{K} D'_{i_{\text{redu}}}(\beta^{(t)}, \alpha^{(t)}, \hat{G}) \left(V_{i_{\text{redu}}}(\beta^{(t)}, \alpha^{(t)}, \hat{G}) \right)^{-1} D_{i_{\text{redu}}}(\beta^{(t)}, \alpha^{(t)}, \hat{G}) \right]^{-1} \\ \times \left[\sum_{i=1}^{K} (D_{i_{\text{redu}}}(\beta^{(t)}, \alpha^{(t)}, \hat{G}))' (V_{i_{\text{redu}}}(\beta^{(t)}, \alpha^{(t)}, \hat{G}))^{-1} S_{i_{\text{redu}}}(\beta^{(t)}, \alpha^{(t)}, \hat{G}) \right]$$
(3.1.21)

3.2 Simulation study

After proposing the second-order generalized estimating equation for mixed models, simulation studies are needed to investigate the finite sample performance of the proposed method in terms of bias and efficiency. Then, compare the simulated outcomes with first order generalized estimating equation for mixed models proposed in Chapter 2. The language **R** has been used for all the generation and the calculation of the simulation data. The **R** codes used for this purpose are available in the appendix. The references of **R** codes are Xu (2013) and Pavlou (2012). We generate the data in the same manner as shown in Chapter 2.

We estimate the parameter coefficients and the variance matrix of the random effect using estimating equation 3.1.11. Then, we compare the estimates and the standard deviations to the estimates and the standard deviations in Chapter 2. Then, we estimate the fixed effects parameter and the correlation parameter α using the estimating equation 3.1.20. The results of these estimation are shown in Tables 3.1 - 3.4.

Table 3.1: Parameter estimates (standard deviations in parentheses) of the true beta values (3.00, 0.50, 1.00, 0.20) when number of clusters K = 50, 100 and 150 with $n_i = 4$ per each cluster, is compared for: the proposed methods (GEEM) in Chapter 2, and the second-order GEEM in Chapter 3, for 5,000 simulations applying the model 2.4.1 with correlation matrix as shown in 2.4.13

K	Parameters	GEEM	GEEM2
	β_0	4.360(.99841)	4.357(.999839)
	β_1	0.654(.00529)	0.661(.009265)
K=50	β_2	0.664(.02157)	0.671(.034598)
	β_3	0.222(.00540)	0.213(.010155)
	$\operatorname{var}(b_{0i})$	—	0.528
	$\operatorname{var}(b_{1i})$	_	0.301
	eta_0	4.225(.99930)	4.1820(.977966)
	β_1	0.691(.00763)	0.6857(.006670)
K=100	β_2	0.912(.02050)	0.8884(.018058)
	β_3	0.187(.00893)	.1956 (.0039883)
	$\operatorname{var}(b_{0i})$	—	0.502
	$\operatorname{var}(b_{1i})$	_	0.306
	eta_0	4.242(.97899)	4.177(1.187227)
K=150	β_1	0.686(.00367)	0.683(.006486)
	β_2	0.925(.01733)	0.898(.016719)
	eta_3	0.197(.00459)	0.198(.004231)
	$\operatorname{var}(b_{0i})$	—	0.469
	$\operatorname{var}(b_{1i})$	_	0.319

Table 3.2: Parameter estimates (standard deviations in parentheses) of the true beta values (3.00, 0.50, 1.00, 0.20) when number of clusters K = 50, 100 and 150 with $n_i = 4$ per each cluster, is compared for: the proposed methods (GEEM) in Chapter 2, and the second-order GEEM in Chapter 3, for 5,000 simulations applying the model 2.4.1 with correlation matrix as shown in 2.4.10

K	Parameters	GEEM	GEEM2
	β_0	4.392(.97769)	4.214(.98296)
K=50	β_1	0.654(.00775)	0.672(.00698)
	β_2	0.665(.02776)	0.742(.02750)
	eta_3	0.213(.00816)	0.199(.00818)
	$\operatorname{var}(b_{0i})$	—	0.375
	$\operatorname{var}(b_{1i})$	_	0.318
	ß	4 254(00784)	4 548(00887)
V 100	ρ_0	4.234(.99704)	4.340(.99001)
K=100	ρ_1	0.085(.00583)	0.692(.00736)
	β_2	0.898(.01843)	0.921(.01884)
	eta_3	0.192(.00560)	0.185(.00502)
	$\operatorname{var}(b_{0i})$	_	0.496
	$\operatorname{var}(b_{1i})$	—	0.305
	Bo	4 269(97485)	4 558(98670)
V -150	β_0	0.681(.00745)	0.600(.00842)
K =130	ρ_1	0.081(.00745)	0.090(.00842)
	eta_2	0.918(.02080)	0.937(.01035)
	eta_3	0.198(.00675)	0.192(.00748)
	$\operatorname{var}(b_{0i})$	—	0.457
	$\operatorname{var}(b_{1i})$	_	0.302

Table 3.3: Parameter estimates (standard deviations in parentheses) of the true beta values (3.00, 0.50, 1.00, 0.20) when number of clusters K = 50, 100 and 150 with $n_i = 4$ per each cluster, is compared for: the proposed methods (GEEM) in Chapter 2, and the second-order GEEM in Chapter 3, for 5,000 simulations applying the model 2.4.1 with correlation matrix as shown in 2.4.9

K	Parameters	GEEM	GEEM2
	eta_0	4.407(.97396)	4.247(.98469)
K=50	β_1	0.647(.00752)	0.666(.00766)
	β_2	0.655(.02290)	0.625(.02339)
	β_3	0.217(.00694)	0.222(.00754)
	α	—	0.837
	eta_0	4.265(1.00192)	4.575(1.00232)
K=100	β_1	0.665(.00578)	0.685(.00602)
	β_2	0.889(.01677)	0.897(.01620)
	β_3	0.194(.00557)	0.192(.00563)
	α	—	0.877
	eta_0	4.259(.97736)	4.587(.98362)
K=150	β_1	0.683(.00474)	0.684(.00523)
	β_2	0.921(.01794)	0.920(.01946)
	β_3	0.197(.00442)	0.196(.00504)
	α	_	0.857

Table 3.4: Parameter estimates (standard deviations in parentheses) of the true beta values (3.00, 0.50, 1.00, 0.20) when number of clusters K = 50, 100 and 150 with $n_i = 4$ per each cluster, is compared for: the proposed methods (GEEM) in Chapter 2, and the second-order GEEM in Chapter 3, for 5,000 simulations applying the model 2.4.1 with correlation matrix as shown in 2.4.12

K	Parameters	GEEM	GEEM2
	β_0	4.352(.98275)	4.609(.9850)
K=50	β_1	0.631(.00925)	0.654(.01017)
	β_2	0.674(.01779)	0.741(.0234)
	eta_3	0.241(.00866)	0.218(.0114)
	lpha	—	0.345
	eta_0	4.223(.99228)	4.551(.99351)
K=100	β_1	0.682(.00573)	0.684(.00574)
	β_2	0.909(.01922)	0.914(.01495)
	eta_3	0.197(.00702)	0.192(.00807)
	α	—	0.373
	eta_0	4.252(.97341)	4.567(.98523)
K=150	β_1	0.678(.00670)	0.680(.00721)
	β_2	0.920(.01824)	0.912(.02041)
	eta_3	0.202(.00564)	0.199(.00405)
	α	_	0.370

In this chapter we focus our interest either on both β and the variance of the random effects b with regarding the correlation parameter α as nuisance or on the coefficient parameters and the association structure α and regard the variance of the random effects as nuisance. We build additional system of estimating equations analogous to equation 2.2.6 can serve to estimate either β and G, simultaneously or β and α , simultaneously.

When our interest focuses either on both β and G or on both β and α , we report the simulation results using equation 3.1.11 in Tables 3.1- 3.4 for various choices of correlation matrix. We compare the simulation results of second-order generalized estimating equation for mixed models which proposed in this chapter with first-order generalized estimating equations for mixed models which proposed in Chapter 2. Although the estimates produced by GEEM2 are slightly higher than those produced by GEEM method proposed in Chapter 2 but the difference is very small under all samples sizes. For the standard deviations, it can be seen that in almost all cases the GEEM2 produces estimates with higher standard deviations but the differences are very small. In conclusion, Both methods perform well across the simulations experiments and choosing between these two methods depends on the goal of the study.
CHAPTER 4 REAL DATA APPLICATIONS

4.1 Introduction

In estimating parameter estimates, standard errors, the random effects and the covariance structures matrices, the simulation studies in Chapter 2 and Chapter 3 have provided quite efficient results in comparison to the other methods.

To further investigate the behavior of the proposed models, the new approaches are applied to real-life data (epilepsy data) that was studied by many others Fitzmaurice et al. (2012). Generalized estimating equations approach is known to provide consistent estimates Liang and Zeger (1986). Therefore, evaluating the performance of the approaches that was proposed in the previous chapters will be done by comparing the parameter estimates and standard errors with those from GEE.

4.2 Data description

According to Fitzmaurice et al. (2012), in 1987 the data was collected on 59 epileptics in placebo-controlled clinical under rigorous controls Leppik et al. (1987). Subjects with repeated seizures were registered in a randomized clinical trial for the treatment of epilepsy. The patients were randomized to either the anti-epileptic drug (Progabide) or the placebo. Before the treatment started, the number of seizures were counted up over an 8-week interval for each subject Fitzmaurice et al. (2012). The goal is to examine weather Progabide reduces the number of seizures significantly compared to placebo Berridge and Crouchley (2011). See Figure 4.1 for the number of seizures for each subject. Figure 4.2 and Figure 4.3 show the baseline data of number of seizures for each subject in placebo or Progabide group, respectively.



Figure 4.1: Number of seizures per each subject in the sample during 8-week prior to the treatment.



Figure 4.2: Number of seizures per each subject in the placebo group over 8-week period prior to the treatment.



Figure 4.3: Number of seizures per each subject in Progabide group over 8-week period prior to the treatment.

The data shows a lot of variation in the sense that the number of seizures for some subjects are very extreme in terms of having unusual repeated seizures from other values in the random sample. Either having very high or very low number of seizures compared to other observations of the patients in the random sample.

Subject 3 and subject 4 are examples of patients who recorded very low number of seizures. Subject 3 had 4 seizures during the 8-week period which is the baseline period and continued to have either very low number of seizures or no seizures at all during the visits after the treatment has started, and subject 4 had 8 seizures during the baseline period and continued to have very low numbers during the visits as well.

While subject 49 and subject 18 are examples of patients who recorded very high numbers of seizures, subject 49 had 151 during the same period and continued to have very high number of observations during the visits after the treatment has started but the subject had 111 seizures during the baseline seizure and he recorded low number of seizures during the visits. His/her observations lie abnormal from other observations.

Figure 4.4 and Figure 4.5 show the number of seizures during the baseline period and during the visits post randomization for placebo and Progabide groups, respectively. Also, Figure 4.6 and Figure 4.7 show the boxplot of log number of seizures during the baseline period and during the visits post randomization for placebo and Progabide groups, respectively.

The data was collected to investigate the effectiveness of the treatment (Progabide). The question here is, does the Progabide drug reduce the number of seizures? To answer this question we will perform mixed models of this data using the newly proposed approach in Chapter 2.



Figure 4.4: Boxplots of numbers of seizures for the placebo group during the baseline period and during visits post randomization.



Figure 4.5: Boxplots of numbers of seizures for the Progabide group during the baseline period and during visits post randomization.



Figure 4.6: Boxplots of log of numbers of seizures for the placebo group during the baseline period and during visits post randomization.



Figure 4.7: Boxplots of log of numbers of seizures for the Progabide group during the baseline period and during visits post randomization.

Progabide



Figure 4.8: Number of seizures per each subject in the sample during 8-week prior to the treatment.



Figure 4.9: Number of seizures per each subject in the sample during 8-week prior to the treatment.

Each subject was observed individually by counting the number of seizures for an 8-weak period before randomization. Then, the subjects were randomized to be treated with either placebo or Progabide with observing the number of seizure for a 2-week period during 4-visits. A total of 59 patients with a total of 4-visits, each subject has four observations and in total there are 236 observations after post randomization. Including baseline observation, each patient has 5 observations and in total there are 295 observations. In the placebo group, we have 28 patients with 140 observations and the remaining patients are assigned randomly into Progabide group. There is definitely individual level variation in each group and there is variation between both groups in general. Figure 4.1 shows heterogeneity between the 95 patients.

4.3 A GEEM Model for the Seizure Data

Here, we fit the model by letting the response explained by the same covariates shown below. Table 4.1 shows epilepsy data of 4 participants out of 59.

- **y** : the response that consists of the number of seizures that had happened prior to the start of the experiment and this is in the 8-week period, and the number of seizures post randomization as well.
- **age**: of each subject at the experiment time.
- **Treatment** : the treatment that had been used post randomization which is either placebo or Progabide.

- id : identifying each subject by the number of the measurement. Each subject has five measurements
- **Time**: shows the time period for each subject, 0 represents the baseline measurement, 1 represents the measurement at the first visit after the experiment has started, 2 represents the measurement at the first visit after the experiment has started and so on.
- **W** : shows the prior and post randomization period.

Table 4.1: The data of some participants in epilepsy study. The data was downloaded from website: "http://www.hsph.harvard.edu/fitzmaur/ala2e/epilepsy.sas7bdat"

id	у	age	Treatment	Time	W	log(W)
1	11	31	placebo	0	8	2.08
1	5	31	placebo	1	2	0.693
1	3	31	placebo	2	2	0.693
1	3	31	placebo	3	2	0.693
1	3	31	placebo	4	2	0.693
2	11	30	placebo	0	8	2.08
2	3	30	placebo	1	2	0.693
2	5	30	placebo	2	2	0.693
2	3	30	placebo	3	2	0.693
2	3	30	placebo	4	2	0.693
4	6	36	placebo	0	8	2.08
4	2	36	placebo	1	2	0.693
4	4	36	placebo	2	2	0.693
4	0	36	placebo	3	2	0.693
4	5	36	placebo	4	2	0.693
49	151	22	Progabide	0	8	2.08
49	102	22	Progabide	1	2	0.693
49	65	22	Progabide	2	2	0.693
49	72	22	Progabide	3	2	0.693
49	63	22	Progabide	4	2	0.693
59	12	37	Progabide	0	8	2.08
59	1	37	Progabide	1	2	0.693
59	4	37	Progabide	2	2	0.693
59	3	37	Progabide	3	2	0.693
59	2	37	Progabide	4	2	0.693

70

 y_{ij} = number of seizures for the *i*-th subject at *j*-th measurement.

$$Time_{ij} = \begin{cases} 1 & \text{if } j = 1, 2, 3, 4 \\ 0 & \text{if } j = 0. \end{cases}$$
$$Treatment_i = \begin{cases} 1 & \text{if Progabide} \\ 0 & \text{if placebo.} \end{cases}$$
$$W_{ij} = \begin{cases} 8 & \text{if } j = 0 \\ 2 & \text{if } j = 1, 2, 3, 4. \end{cases}$$

with $i = 1, \dots, 59$ and j = 0, 1, 2, 3, 4. Consider that the conditional model fulfills

$$\log\left(\frac{\mathbf{E}(y_{ij}|b_i)}{W_{ij}}\right) = \beta_0 + \beta_1 Time_{ij} + \beta_2 Treatment_i + \beta_3 Time_{ij} * Treatment_i + b_{0i} + b_{1i} Time_{ij}$$
(4.3.1)

where this model can be rewritten for each treatment group as below. For placebo group 4.3.1 can be written as

$$\log\left(\frac{\mathbf{E}(y_{ij}|b_i)}{W_{ij}}\right) = \begin{cases} \beta_0 + b_{0i} & \text{if } j = 0\\ \beta_0 + b_{0i} + \beta_1 + b_{1i} & \text{if } j = 1, 2, 3, 4. \end{cases}$$
(4.3.2)

and for Progabide group 4.3.1 can be written as

$$\log\left(\frac{\mathbf{E}(y_{ij}|b_i)}{W_{ij}}\right) = \begin{cases} \beta_0 + b_{0i} + \beta_2 & \text{if } j = 0\\ \beta_0 + b_{0i} + \beta_1 + b_{1i} + \beta_2 + \beta_3 & \text{if } j = 1, 2, 3, 4. \end{cases}$$
(4.3.3)

Fitzmaurice et al. (2012). Here, the parameter β_3 is the parameter of interest since the exponen-

tiated of β_3 represents the expected change of rate of seizures for a participant assigned to Progabide treatment compared to a participant assigned to placebo treatment in post randomization period Wakefield (2009). The response y_{ij} is explained by the time, treatment in addition to the interaction between the time and treatment.

Note that, due to the differences in length of the time periods before and after randomization $log(W_{ij})$ is included in the model Fitzmaurice et al. (2012). The random effects

$$b_i \sim N(0, G)$$

where G is 2×2 covariance matrix. The covariance matrix $R_i(\alpha)$ is assumed to be exchangeable - namely, that

$$\operatorname{Corr}(Y_{ij}, Y_{ik}) = \alpha, \ \forall \ j \neq k$$

The big feature of this correlation matrix is that, only one parameter needs to be estimated but it ignores the time varying between observations. It can be written as

$$R_{i,j} = \begin{cases} 1 & i = j \\ \alpha & \text{otherwise} \end{cases}$$

Or in matrix notation for *i*-th subject,

$$R_{i} = \begin{bmatrix} 1 & \alpha & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha & \alpha \\ \alpha & \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & \alpha & 1 \end{bmatrix}$$

The results given in Table 4.2 shows the parameter estimates, standard errors and P-value for the fixed effects part of our model which is what we are interested in. No subject was excluded from the sample, even those who recorded abnormal observations. Deciding whether the Progabide drug is decreasing the number of seizures. The test stated as

$$H_0:\beta_3=0 \quad vs \quad H_1:\beta_3\neq 0$$

Table 4.2 shows that, for 0.05 level the evidence against the null hypothesis is strong, so we reject the null hypothesis and conclude that a significant interaction exists.

For a subject from placebo group, the seizure rate doesn't change after being treated with placebo treatment. On the contrary, the seizure rate of a subject that participated in Progabide does change which is lower than prior to treatment by about 27%. On the other hand, generalized estimating equations estimates obtained by the **gee()** function in **R Language** which provides slightly different results but they are still very close and indicate the same results in general as shown in Table 4.3 and Table 4.4. The GEE methods indicates that for a subject belonging to placebo group has no change in seizure rate after being treated with placebo but for a subject assigned to Progabide treatment the reduction on the seizure rate is 27%.

In general, both of our proposed approaches in Chapter 2 and generalized estimating equations approach using either **geeglm()** or **gee()** show that, for .05 a significant interaction exists. Since the parameter estimates are unknown, we compare our parameter estimates with those for existing GEE which is known to provide consistent estimates.

Fixed Effects	Parameter Estimate	Standard Error	P-value
Intercept	1.2110	0.1880	0.0000
Time	-0.0010	0.0470	0.4730
Treatment	0.0840	0.2078	0.3740
Time $ imes$ Treatment	-0.3141	0.05764	0.0102

Table 4.2: Parameter estimates for fixed effects part applying the newly proposed GEEM approach on epilepsy data

Visualizing the analysis of epilepsy data assesses the results that we have obtained. The plots 4.10 - Figure 4.13 are obtained by using sjplot package. First, with the function sjp.lmer

Table 4.3: Parameter	estimates for applying th	e generalized estimating equation method Liang and
Zeger (1986) using	geeglm() function in	R language

Fixed Effects	Parameter Estimate	Standard Error	P-value
Intercept	1.2124	0.1663	0.0000
Time	- 0.0022	0.04851	0.5600
Treatment	0.0831	0.2690	0.3200
Time $ imes$ Treatment	-0.3200	0.0640	0.0235

we specified **sort.est = "Time"** and then with function **sort.est = "Time"** we

sort the effects in order.

Table 4.4: Parameter estimates for applying the generalized estimating equation method Liang and Zeger (1986) using **gee()** function in **R language**

Fixed Effects	Parameter Estimate	Standard Error	P-value
Intercept	1.29007	0.1658	0.0000
Time	-0.00176	0.0482	0.366
Treatment	0.03778	0.2087	0.1811
Time $ imes$ Treatment	-0.25301	0.0575	0.0095

For the placebo group, Figure 4.10 and Figure 4.11 show that there are quite some variations in the intercept. While for **Time** there is not much of a variation. It is more or less around the specific line. There are quite some variations in the intercept for subjects who participated in the Progabide group and for this group **Time** shows slight variation.



Figure 4.10: The plot of the random effects of random intercept and random coefficient **Time** for placebo group.



Figure 4.11: Resorting the plot of the random effects of random intercept and random coefficient **Time** for placebo group.



Figure 4.12: The plot of the random effects of random intercept and random coefficient **Time** for Progabide group.



Figure 4.13: Resorting the random effects of random intercept and random coefficient **Time** for Progabide group.

4.4 A GEEM2 Model for Epilepsy Data

Deciding whether the Progabide drug is decreasing the number of seizures using GEEM2 approach. The test stated as

$$H_0:\beta_3=0 \quad vs \quad H_1:\beta_3\neq 0$$

Table 4.5 and 4.6 show that, for 0.05 level the evidence against the null hypothesis are strong, so we reject the null hypothesis and conclude that a significant interaction exists.

For a subject from the placebo group, the seizure rate doesn't change after being treated with the placebo treatment. On the contrary, the seizure rate of a subject that participated in Progabide does change which is lower than prior to treatment by about 27%.

Table 4.5:	Parameter	estimates	for fixed	effects	part and	l variance	of random	effect	applying	the
proposed	GEEM2 app	proach on	epilepsy o	lata.						

Fixed Effects	Parameter Estimate	Standard Error	P-value
Intercept	1.2120	0.1800	0.0000
Time	-0.0020	0.0410	0.4850
Treatment	0.0840	0.2078	0.3740
Time \times Treatment	-0.3071	0.05764	0.0102
Var(b_{0i})	0.527	_	_
Var(b_{1i})	0.201	_	—

For the comparison between the two newly proposed methods, it can be seen that both preform very well. The estimates of β are very closed in the two methods. The first-order of the generalized estimating equation slightly produces lower standard errors.

Tables 4.2, 4.5 and 4.6 show that in both methods the seizure rate doesn't change after being treated with the placebo treatment while the seizure rate of a subject that participated in Progabide does change which is lower than prior to treatment by about 27% in GEEM and lower than prior to treatment by about 26% in GEEM2.

Table 4.6: Parameter estimates for fixed effects part and the correlation parameter applying the proposed GEEM2 approach on epilepsy data.

Fixed Effects	Parameter Estimate	Standard Error	P-value
Intercept	1.1100	0.1800	0.0000
Time	0.0018	0.0470	0.4770
Treatment	0.0840	0.2667	0.3010
Time \times Treatment	-0.3110	0.0518	0.0090
α	0.817	_	_

CHAPTER 5 SUMMARY AND CONCLUSION

Generalized linear mixed model consisted of two parts, fixed effects and random effects. This approach is very helpful and widely used in different fields. It differs than linear mixed models by so unique features such as: It is applicable to any distribution that belongs to the exponential family not only normal distribution and in case of abnormality appropriate link function is often applied.

Generalized estimating equation is a simple method since it does not depend on the likelihood which is not plausible due to its difficulty of maximization. This method has a unique feature as it is known to provide consistent estimates.

5.1 Conclusion Remarks

In this dissertation, we combined the above two methods and introduced first-order of generalized estimating equations for mixed models and then extended it to second-order. The newly proposed approaches in Chapter 2 and its extension in Chapter 3 are general approaches based on continuous and discrete distributions in exponential family. We mainly focused on the longitudinal data generated from Poisson distribution with normal random effects.

In the first-order generalized estimating equations for mixed model, we primarily focused on the fixed effects. We regard the association parameter and the variance of the random effect as nuisances in Chapter 2. We proposed estimating equation to estimate the regression coefficients β with extracting the values of the variance of the random effects *G* and the association parameter α from well known methods such as method of moment for the association parameter and GLMM method for the variance of the random effects. These estimating equations methods have no closed form solution and we solved iteratively using Newton-Raphson as it is a popular iterative method for generalized linear model and generalized estimating equations. The second-order generalized estimating questions for mixed model in Chapter 3 is an extension of the first-order in Chapter 2. We proposed additional estimating equations as an analogue to the essential estimating equations to have a parallel system. This new system can estimate additional parameter in addition to the coefficient regression parameters β . First, we primarily focused on the fixed effects and the variance of the random effects. We regard the association parameter as a nuisance with extracting its value using method of moments. Second, we focus on the fixed effect and the association parameter and regard the variance of the random effects as a nuisance. We extended Newton-Raphson to handle this parallel system.

A simulation study was conducted to investigate the behaviors of the newly proposed methods, in section 2.4 for GEEM and in section 3.2 for GEEM2. In section 2.4, we compare the first-order generalized estimating for mixed models approach with the results of some selected methods. The simulation study has been done under various correlation matrices ranging between strong correlation to uncorrelated data. The results reported in Tables 2.1-2.4 show that the estimates of β are very close to the estimates which are provided by GEE method more than any other method. Moreover, our proposed method has reduced the standard deviations of the GEE method Liang and Zeger (1986)

Furthermore, in section 3.2 we conducted extensive simulation studies and compared the results of our proposed method in Chapter 2 to its extension in Chapter 3. The proposed method performs very well in reducing the standard errors as the samples size increases. It is worth pointing out that as the sample size K increases the difference between the standard errors of our proposed methods decreases.

We finally apply the proposed method to real-life data (epilepsy data) to further evaluate its behavior. The result shows that the GEEM and GEEM2 methods provide consistent estimates. In general, this kind of estimation equations have been shown and proven by many authors to provide consistent estimates. Since the correlation of the data is unknown and this might lead to misspecification of the model and hence the standard errors might get effected. Therefore, the sandwich method have been used to adjust the standard errors.

5.2 Future Research

In the future, we plan to extend our proposed approach to analyze the correlated data with measurement error. We assume the collected data are accurate while it is not due to measurement error that may arise from various sources which can lead to severe bias. A measurement error takes place if we cannot exactly observe some variables, either one or more, in the underlying model. We need to correct the measurement error to obtain unbiased estimates for the parameters of interest. According to Buonaccorsi (2010), a measurement error model with LMM can be expressed as

$$X_i\beta + Z_ib_i = X_{i1}\beta_1 + X_{i2}\beta_2 + Z_{i1}b_{i1} + Z_{i2}b_{i2}, (5.2.1)$$

where X_{i2} and Z_{i2} are observed exactly while X_{i1} and Z_{i1} are subject to measurement error. If we only consider the simpler case that β_1 is a scalar and $Z_{i1} = 0$, i.e. the case that there is no measurement error occurring to the random effect part of the model. Put $Z_{i2} = Z_i$ and $b_{i2} = b_i$, so that

$$Y_i = \beta_1 X_{i1} + X_{i2} \beta_2 + Z_i b_i + \epsilon_i.$$
(5.2.2)

The measurement error for X_{i1} is assumed to be additive:

$$W_i = X_{i1} + u_i, (5.2.3)$$

where W_i is the error-prone measure of X_{i1} , $E(u_i) = 0$, $Cov(u_i) = \Sigma_{ui}$ and u_1, \dots, u_n are assumed independent. In addition to 5.2.2 and 5.2.3 it is assumed that $E(X_{i1}) = \mu_{Xi}$ and $Cov(X_{i1}) = \Sigma_X$. Both u_i in 5.2.3 and X_{i1} are independent of random quantities in 5.2.2. Then,

$$E(W_i) = \mu_{Xi}, Cov(W_i) = \Sigma_W = \Sigma_X + \Sigma_u.$$

To sum up, Y_i, W_i, X_{i2} and Z_i are available data for analysis on the regression parameters β_1, β_2 of interest.

BIBLIOGRAPHY

- Berridge, D. M. and R. Crouchley (2011). *Multivariate generalized linear mixed models using R*. CRC Press.
- Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* 88(421), 9–25.
- Buonaccorsi, J. P. (2010). Measurement error: models, methods, and applications. CRC Press.
- Carey, V., S. L. Zeger, and P. Diggle (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* 80(3), 517–526.
- Chaganty, N. R. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations. *Journal of Statistical Planning and Inference* 63(1), 39–54.
- Chen, Z. (2010). Analysis of Correlated Data with Measurement Error in Responses or Covariates. University of Waterloo.
- Crowder, M. (2001). On repeated measures analysis with misspecified covariance structure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63(1), 55–62.
- Diggle, P. (2002). Analysis of longitudinal data. Oxford University Press.
- Firth, D. (1991). Generalized linear models. Chapter 3 of Statistical Theory and Modelling, DV Hinkley, N. Reid, EJ Snell, eds.
- Fitzmaurice, G. M., N. M. Laird, and J. H. Ware (2012). *Applied longitudinal analysis*, Volume 998. John Wiley & Sons.
- Gourieroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood methods: Theory. *Econometrica: Journal of the Econometric Society*, 681–700.

- Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* 72(358), 320–338.
- Heagerty, P. J. and S. L. Zeger (1996). Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association 91*(435), 1024–1036.
- Hemmerle, W. J. and H. O. Hartley (1973). Computing maximum likelihood estimates for the mixed aov model using the w transformation. *Technometrics* 15(4), 819–831.
- Hutcheson, G. D. and N. Sofroniou (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. Sage.
- Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics*, 963–974.
- Lane, S. (2007). *Generalized Estimating Equations for Pedigree Analysis*. Department of Mathematics and Statistics: University of Melbourne.
- Leppik, R. et al. (1987). A controlled study of progabide in partial seizures methodology and results. *Neurology* 37(6), 963–963.
- Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73(1), 13–22.
- Liang, K.-Y., S. L. Zeger, and B. Qaqish (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 3–40.
- Lindstrom, M. J. and D. M. Bates (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, 673–687.
- Lo, C. H., W. K. Fung, and Z. Y. Zhu (2007). Structural parameter estimation using generalized estimating equations for regression credibility models. *ASTIN Bulletin: The Journal of the IAA 37*(2), 323–343.

- McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research 16*(3), 285–292.
- McCullagh, P. and J. Nelder (1983). Generalized Linear Models. Chapman and Hall London.
- McCullagh, P. and J. Nelder (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.
- Molenberghs, G. and G. Verbeke (2005). *Models for discrete longitudinal data*. Springer-Verlag New York.
- Molenberghs, G., G. Verbeke, and C. G. Demétrio (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime data analysis 13*(4), 513–531.
- Nelder, J. A. and R. J. Baker (1972). Generalized linear models. Wiley Online Library.
- Neuhaus, J. M. and C. E. McCulloch (2006). Separating between-and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(5), 859–872.
- Pankhurst, Q. A., J. Connolly, S. Jones, and J. Dobson (2003). Applications of magnetic nanoparticles in biomedicine. *Journal of physics D: Applied physics 36*(13), R167.
- Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58(3), 545–554.
- Pavlou, M. (2012). Analysis of clustered data when the cluster size is informative. Ph. D. thesis, UCL (University College London).
- Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 1033–1048.
- Prentice, R. L. and L. Zhao (1991). Estimating equations for parameter in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 825–839.

Rudary, M. R. (2009). On predictive linear gaussian models. Ph. D. thesis, University of Michigan.

- Sutradhar, B. C. and V. Jowaheer (2003). On familial longitudinal poisson mixed models with gamma random effects. *Journal of multivariate analysis* 87(2), 398–412.
- Swan, T. (2006). Generalized estimating equations when the response variable has a Tweedie distribution: An application for multi-site rainfall modelling. Ph. D. thesis, University of Southern Queensland.
- Wakefield, J. (2009). Stat/Biostat 571 Statistical Methodology Regression Models for dependent data. http://courses.washington.edu/b571/lectures/set1.pdf.
- Wang, X., S. Lee, X. Zhu, S. Redline, and X. Lin (2013). Gee-based snp set association test for continuous and discrete traits in family-based association studies. *Genetic epidemiology* 37(8), 778–786.
- Xu, S. (2013). Generalized estimating equation based zero-inflated models with application to examining the relationship between dental caries and fluoride exposures. University of Louisville.
- Yi, G. Y. and R. J. Cook (2002). Marginal methods for incomplete longitudinal data arising in clusters. *Journal of the American Statistical Association* 97(460), 1071–1080.
- Zeger, S. L. and K.-Y. Liang (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 121–130.
- Zhao, L. P. and R. L. Prentice (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* 77(3), 642–648.
- Zhao, L. P., R. L. Prentice, and S. G. Self (1992). Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 805–811.
- Ziegler, A., C. Kastner, and M. Blettner (1998). The generalised estimating equations: an annotated bibliography. *Biometrical Journal 40*(2), 115–139.

Ziegler, A., C. Kastner, U. Grömping, and M. Blettner (1996). The generalized estimating equations in the past ten years: An overview and a biomedical application.

APPENDIX A SELECTED R PROGRAMS

• A sample of complete code that gives the result shown in Table 2.1 when K = 50.

##Install Packages #install.packages(gee) #install.packages(Matrix) #install.packages(lme4) #install.packages(MASS) #install.packages(corcounts) #install.packages(mmm) #install.packages(geepack) #install.packages(plyr) #install.packages(dplyr) #install.packages(haven) #install.packages(tidyr) #install.packages(dplyr) #install.packages(glmmML) #install.packages(geepack) #install.packages(brms) #install.packages(mvtnorm)

##Load packages
library(gee)
library(Matrix)
library(lme4)
library(MASS)

```
##Generate correlated count data

#sim.long.count1 <-function(seed){

set.seed(1)

k <- 50 ;T.<-4;

beta0 <-3

beta1 <-.5

beta2 <-1

beta3 <-.2

n<-rep(4,k)

n1=16
```

```
z1=matrix (1, sum(n), 1)
y1=matrix (0, sum(n), 1)
x1=matrix (0:3, sum(n), 1)
```

```
r11= matrix (0, (sum(n)/2) - n1, 1)
r12= matrix (1, (sum(n)/2) + n1, 1)
r1=rbind (r11, r12)
```

```
margins <- c("Poi","Poi","Poi","Poi"); ranef.covar=diag(c(.4,.3));
b.11=rmvnorm(n=1, sigma=ranef.covar)
mu1 <- c(exp(beta0+beta1*x1[1]+beta2*r1[1]+beta3*x1[1]*r1[1]
+b.11[1,1]+b.11[1,2]*x1[1]), exp(beta0+beta1*x1[2]+beta2*r1[2]
+beta3*x1[2]*r1[2]+b.11[1,1]+b.11[1,2]*x1[2]),
exp(beta0+beta1*x1[3]+beta2*r1[3]+beta3*x1[3]*r1[3]+b.11[1,1]
+b.11[1,2]*x1[3]),
```

```
exp (beta0+beta1*x1[4]+beta2*r1[4]+beta3*x1[4]*r1[4]
+b.11[1,1]+b.11[1,2]*x1[4]))
corstr <- "unstr"
corpar1 <- matrix (c(1,0.4,0.6,0.7,
0.4,1,0.6,0.37,
0.6,0.6,1,0.6,
0.7,0.37,0.6,1), ncol=T., byrow=T.)
```

```
Y. begining1 <- rcounts (N=k, margins=margins, mu=mu1, corstr=corstr, corpar=corpar1)
```

```
Y. begining11 <- matrix (c(Y. begining1 [1:(((sum(n)/2)+n1)/T.),1],</p>
Y. begining1 [1:(((sum(n)/2)+n1)/T.),2],
Y. begining1 [1:(((sum(n)/2)+n1)/T.),3],
Y. begining1 [1:(((sum(n)/2)+n1)/T.),4]),((sum(n)/2)+n1)/T., T.)
```

```
 mu2 <- c(exp(beta0+beta1*x1[1]+beta2*r1[(sum(n)/2)+n1] + beta3*x1[1]*r1[(
```

```
sum(n)/2)+n1]+b.11[1,1]+b.11[1,2]*x1[1]),
exp (beta0+beta1*x1[2]+beta2*r1[(sum(n)/2)+n1]
+beta3*x1[2]*r1[(sum(n)/2)+n1]+b.11[1,1]+b.11[1,2]*x1[2]),
exp(beta0+beta1*x1[3]+beta2*r1[(sum(n)/2)+n1]+beta3*x1[3]*r1[(sum(n)/2)+n1]+b.11[1,1]+b.11[1,2]*x1[3]),
exp (beta0+beta1*x1[4]+beta2*r1[(sum(n)/2)+n1]+
beta3*x1[4]*r1[(sum(n)/2)+n1]+b.11[1,1]+b.11[1,2]*x1[4]))
```

```
Y. begining2 <- rcounts (N=k, margins=margins, mu=mu2, corstr=corstr, corpar=corpar1)</p>
Y. begining12 <- matrix (c(Y. begining2 [((((sum(n)/2)+n1)/T.)+1):k, 1], Y. begining2 [((((sum(n)/2)+n1)/T.)+1):k, 2], Y. begining2 [((((sum(n)/2)+n1)/T.)+1):k, 3], Y. begining2 [((((sum(n)/2)+n1)/T.)+1):k, 4]), ((sum(n)/2)-n1)/T., T. ))</p>
```

```
Y. begining1 <- rbind(Y. begining11, Y. begining12)
id1<-matrix(rep(seq(1:k),4))
ID=as.matrix(rep(1:k,n), n,1)
dat1=list( x1=x1, y1=y1, z1=z1, ID=ID, r1=r1)
# return(dat1)
```

sim.long.count1<-as.data.frame(cbind(ID,y1,x1,z1,r1))
names(sim.long.count1)<-c("ID","y1","x1","z1", "r1")</pre>

```
GEE. Mixed. Simu<-function (sim.long.count1){
dat1<-data.frame(id = sample(sim.long.count1$ID,k*T.,rep=F))
```

```
countss <- sim.long.count1 %>%
group_by(ID) %>%
do(data.frame(nrow=nrow(.)))
```

#Split the data into the clusters. #Each sbject is defined as a "list" cluster <--list() idua <--unique(sim.long.count1\$ID) for (i in 1:length(idua)) cluster[[i]]<--sim.long.count1[sim.long.count1\$ID==idua[i],]</pre>

y<-sim.long.count1\$y1
x<-as.matrix(sim.long.count1\$x1)
r<-sim.long.count1\$r1
z<-as.matrix(sim.long.count1\$x1)
#Fit a geeglm to obtain a vector of initial estimates for beta
(betain)</pre>

C.M1 = 5000

beta00cm1=numeric (C.M1) beta11cm1=numeric (C.M1) beta22cm1=numeric (C.M1) beta33cm1=numeric (C.M1)

for(1 in 1:C.M1){

Y. begining1 <- rcounts (N=k, margins=margins, mu=mul, corstr=corstr,

- Y. begining 11 < -matrix (c(Y. begining 1 [1:(((sum(n)/2)+n1)/T.), 1]),
- Y. begining1 [1:(((sum(n)/2)+n1)/T.), 2],
- Y. begining1 [1:(((sum(n)/2)+n1)/T.),3],
- Y. begining1 [1:(((sum(n)/2)+n1)/T.), 4]), ((sum(n)/2)+n1)/T., T.)

```
Y. begining2 <- rcounts (N=k, margins=margins, mu=mu2, corstr=corstr, corpar=corpar1)</p>
Y. begining12 <- matrix (c(Y. begining2 [((((sum(n)/2)+n1)/T.)+1):k, 2], k, 1], Y. begining2 [((((sum(n)/2)+n1)/T.)+1):k, 2], Y. begining2 [((((sum(n)/2)+n1)/T.)+1):k, 3], Y. begining2 [((((sum(n)/2)+n1)/T.)+1):k, 4]), ((sum(n)/2)-n1)/T., T. )</p>
```

```
Y. begining1 <- rbind (Y. begining11, Y. begining12)
```

```
y1<-y1+as.vector(t(Y.begining1))
```

```
geeglm.simu <- geeglm(formula = y1 ~ x1 + r1 + x1:r1,
family = poisson(link = "log"),
id = ID,
corstr = "ar1",
scale.fix = FALSE)
betain0= geeglm.simu$coefficients[1]
```

```
betain1= geeglm.simu$coefficients[2]
```

```
betain2= geeglm.simu$coefficients[3]
betain3= geeglm.simu$coefficients[4]
betain=c(betain0, betain1, betain2, betain3)
```

```
#N is the number of clusters
N<-length (countss$ID)
#p is the number of parameters
p<-length (betain)
accuracy <- 0.001
error <-1
#itera is the number of iterations
itera <-0
while (error > accuracy){
I0 < -0; I1 < -0; I2 < -0
itera <-- itera +1
for(i in 1:length(idua)){
#ni is the number of members in the ith cluster
ni<-length(cluster[[i]]$ID)
x. full \ll cbind(1, cluster[[i]] x,
cluster [[i]] $r, cluster [[i]] $x*cluster [[i]] $r)
z.full \ll cbind(1, cluster[[i]] x)
y<-as.vector(cluster[[i]]$y)
```

```
G<-matrix(c(gg$vcov[1],0,0, gg$vcov[2]),nrow=2, ncol = 2)

mu<-list()

j=0

for (i in 1:nrow(x.full)){

mu[[i]] <- as.vector(x.full[i, ]%*%betain +
```

```
D<-M%*%x.full
```

```
H0.beta <- matrix (rep(0,p*p), nrow=p, ncol=p)
H1.beta <- matrix (rep(0,p), nrow=p, ncol=1)
```

```
Ri. Matrix <-- matrix (rho, nrow=ni, ncol=ni)
diag (Ri. Matrix)<-1
```

```
T.<−4
J <- matrix (c(1), T., T.)
Vi. Matrix <-M%*%exp(z.full%*%G%*%t(z.full)-J)%*%M+
(M)%*%Ri. Matrix
```

```
H0.beta <-H0.beta+t(D)%*% ginv(Vi.Matrix)%*%D
H1.beta <- H1.beta+t(D)%*% ginv(Vi.Matrix) %*%((y-mu))
```

```
I0in <-(t (D)%*%ginv (Vi. Matrix)%*%D)
I0 <-(I0+I0in)
```

```
Ilin <-t (D)%*%ginv(Vi. Matrix)%*%(y-mu)
Il <-Il+Ilin
I2in <-t (D)%*%ginv(Vi. Matrix)%*%(y-mu)%*%t (y-mu)%*%
ginv(Vi. Matrix)%*%D
I2 <-I2+I2in
}
#sigma2.delta <-(y-mu.new)^2/(N-p)
beta.new<- (betain + ginv(I2) %*% H1.beta)
for (i in 1:nrow(x.full)){
mu.new <- as.vector(x.full[i,]%*%beta.new +
t(z.full[i,])%*%G%*%z.full[i,])</pre>
```

}

```
#rho <-((y[i]-mu.new)%*%t(y[i]-mu.new))
rho <- 0
error <-sum((beta.new-betain)^2)
betain <-beta.new
mu <-mu.new
if (!(itera <25)) print("Iterations > 25")
if (!(itera <25))
return(list(Converge="Error"))
}</pre>
```

```
beta00cm1 [[1]]= beta . new [1]
beta11cm1 [[1]]= beta . new [2]
beta22cm1 [[1]]= beta . new [3]
beta33cm1 [[1]]= beta . new [4]
```

```
robust <-(ginv(I0)%*%(I2)%*%ginv(I0))
beta.GEEM.cm1=c(mean(beta00cm1), mean(beta11cm1),
mean(beta22cm1), mean(beta33cm1))
sd.cm1=c(sd(beta00cm1), sd(beta11cm1), sd(beta22cm1),
sd(beta33cm1))</pre>
```

```
s.e.cml=c(sd[1]/sqrt(C.M1), sqrt(sd[2]/(C.M1)), sd[3]/sqrt(C.M1),
sd[4]/sqrt(C.M1))
se.beta<-sqrt(((diag(robust))[1:p]))/N
beta.fit.cml<-cbind(beta=beta.new, S.E=(se.beta),
,M.C.Estimate=beta.GEEM.cml, M.C.S.E=(s.e.cml),
M.C.S.D=(sd.cml))
colnames(beta.fit.cml)<-c("Estimate", "S.E.",
"M.C.Estimate", "M.C.S.E", "M.C.S.D")
Result.cml=list(Converge="YES", Beta=round(beta.fit.cml
,4), Number_of_Iterations=itera)
return(Result.cml)
}
GEE.Mixed.Simu( sim.long.count1)
```

• A sample of complete code that gives the result shown in Table 3.1 when K = 50.

```
##Generate correlated count data
set.seed(1)
k <- 50; T. < -4;
beta0 <-3
beta1 <-.5
beta2 <-1
beta3 <-.2
n <-rep(4,k)
```
z1 = matrix(1, sum(n), 1)

```
y_1 = matrix(0, sum(n), 1)
x1 = matrix(0:3, sum(n), 1)
r11 = matrix(0, (sum(n)/2) - n1, 1)
r12 = matrix(1, (sum(n)/2) + n1, 1)
r1 = rbind(r11, r12)
margins <- c("Poi","Poi","Poi","Poi")
ranef.covar=diag(c(.4,.3))
b.11=rmvnorm(n=1, sigma=ranef.covar)
mu1 < -c(exp(beta0+beta1*x1[1]+beta2*r1[1]+beta3*x1[1]*r1[1])
+b.11[1,1]+b.11[1,2]*x1[1]),
\exp(beta0+beta1*x1[2]+beta2*r1[2]+beta3*x1[2]*r1[2]
+b.11[1,1]+b.11[1,2]*x1[2]),
exp(beta0+beta1*x1[3]+beta2*r1[3]+beta3*x1[3]*r1[3]
+b.11[1,1]+b.11[1,2]*x1[3]),
exp (beta0+beta1*x1[4]+beta2*r1[4]+beta3*x1[4]*r1[4]
+b.11[1,1]+b.11[1,2]*x1[4]))
corstr <- "unstr"
corpar1 <- matrix (c(1,0.4,0.6,0.7,
0.4,1,0.6,0.37,
0.6, 0.6, 1, 0.6,
0.7, 0.37, 0.6, 1, ncol=T., byrow=T.)
Y. begining1 <- rcounts (N=k, margins=margins, mu=mul, corstr=corstr,
corpar=corpar1)
```

Y. begining 11 < -matrix (c(Y. begining 1 [1:(((sum(n)/2)+n1)/T.), 1]),

```
Y. begining1 [1:(((sum(n)/2)+n1)/T.),2],
Y. begining1 [1:(((sum(n)/2)+n1)/T.),3],
Y. begining1 [1:(((sum(n)/2)+n1)/T.),4]),((sum(n)/2)+n1)/T., T.)
```

```
\begin{split} & mu2 <- c(exp(beta0+beta1*x1[1]+beta2*r1[(sum(n)/2)+n1] \\ &+beta3*x1[1]*r1[(sum(n)/2)+n1]+b.11[1,1]+b.11[1,2]*x1[1]), \\ & exp(beta0+beta1*x1[2]+beta2*r1[(sum(n)/2)+n1] \\ &+beta3*x1[2]*r1[(sum(n)/2)+n1]+b.11[1,1]+b.11[1,2]*x1[2]), \\ & exp(beta0+beta1*x1[3]+beta2*r1[(sum(n)/2)+n1]+ \\ & beta3*x1[3]*r1[(sum(n)/2)+n1]+b.11[1,1]+b.11[1,2]*x1[3]), \\ & exp(beta0+beta1*x1[4]+beta2*r1[(sum(n)/2)+n1] \\ &+beta3*x1[4]*r1[(sum(n)/2)+n1]+b.11[1,1]+b.11[1,2]*x1[4])) \end{split}
```

```
Y. begining2 <- rcounts (N=k, margins=margins, mu=mu2, corstr=corstr, corpar=corpar1)</p>
Y. begining12 <- matrix (c(Y. begining2 [((((sum(n)/2)+n1)/T.)+1):k, 2], +1):k, 1], Y. begining2 [((((sum(n)/2)+n1)/T.)+1):k, 2], Y. begining2 [((((sum(n)/2)+n1)/T.)+1):k, 3], Y. begining2 [((((sum(n)/2)+n1)/T.)+1):k, 4]), ((sum(n)/2)-n1)/T., T. )</p>
```

Y. begining1 <- rbind (Y. begining11, Y. begining12)

```
y1<-y1+as.vector(t(Y.begining1))
id1<-matrix(rep(seq(1:k),4))
options( warn = -1 )
ID=as.matrix(rep(1:k,n), n,1)
dat1=list( x1=x1, y1=y1, z1=z1, ID=ID, r1=r1)
sim.long.count1<-as.data.frame(cbind(ID,y1,x1,z1,r1))
names(sim.long.count1)<-c("ID","y1","x1","z1", "r1")</pre>
```

GEE. Mixed. Simu2<-function(sim.long.count1){

```
dat1 <-- data.frame(id = sample(sim.long.count1$ID,k*T.,rep=F))
```

```
countss <- sim.long.count1 %>%
group_by(ID) %>%
do(data.frame(nrow=nrow(.)))
```

```
#Split the data into the clusters.
#Each sbject is defined as a "list"
cluster <-list()
idua<--unique(sim.long.count1$ID)
for (i in 1:length(idua))
cluster[[i]]<--sim.long.count1[sim.long.count1$ID==idua[i],]</pre>
```

```
y<-sim.long.count1$y1
x<-as.matrix(sim.long.count1$x1)
r<-sim.long.count1$r1
z<-as.matrix(sim.long.count1$x1)</pre>
```

M.C=5000

```
beta00.GEEM2G=numeric (M.C)
beta11.GEEM2G=numeric (M.C)
beta22.GEEM2G=numeric (M.C)
beta33.GEEM2G=numeric (M.C)
var.b0=numeric (M.C)
var.b1=numeric (M.C)
```

for (1 in 1:M.C)

Y. begining1 <- rcounts (N=k, margins=margins, mu=mu1, corstr=corstr, corpar=corpar1)

Y. begining 11 < -matrix (c(Y, begining 1 [1:(((sum(n)/2)+n1)/T.), 1]),

Y. begining1 [1:(((sum(n)/2)+n1)/T.), 2],

Y. begining1 [1:(((sum(n)/2)+n1)/T.),3],

Y. begining1 [1:(((sum(n)/2)+n1)/T.), 4]), ((sum(n)/2)+n1)/T., T.)

```
Y. begining2 <- rcounts (N=k, margins=margins, mu=mu2, corstr=corstr, corpar=corpar1)</p>
Y. begining12<-matrix (c(Y. begining2 [((((sum(n)/2)+n1)/T.)+1):k,2], +n1)/T.)+1):k,1], Y. begining2 [((((sum(n)/2)+n1)/T.)+1):k,2], Y. begining2 [((((sum(n)/2)+n1)/T.)+1):k,3], Y. begining2 [((((sum(n)/2)+n1)/T.)+1):k,4]),((sum(n)/2)-n1)/T., T. )</p>
```

Y. begining1 <- rbind (Y. begining11, Y. begining12)

```
y1<-y1+as.vector(t(Y.begining1))
```

```
glmerLaplace.simu <- glmmML(formula = y1 ~ x1 + r1 + x1:r1,
family = poisson, cluster = ID)
```

```
beta0GEEM2G= glmerLaplace.simu$coefficients[1]
beta1GEEM2G= glmerLaplace.simu$coefficients[2]
beta2GEEM2G= glmerLaplace.simu$coefficients[3]
```

```
beta3GEEM2G= glmerLaplace.simu$coefficients[4]
betainGEEMG=c(beta0GEEM2G, beta1GEEM2G, beta2GEEM2G,
beta3GEEM2G)
#Initial variance of random effects
glmerLaplace.simu <- glmer(formula = y1 ~ x1 + r1 + x1:r1 + ( x1 | ID),
data = sim.long.count1,
family = poisson(link = "log"),
nAGQ = 1)
g<-as.data.frame(VarCorr(glmerLaplace.simu))
G<-matrix(c(gg$vcov[1],0,0, gg$vcov[2]),nrow=2, ncol = 2)
rand.eff.b<-matrix(c(gg$vcov[1], gg$vcov[2]),nrow=2, ncol = 1)
U0=c(betainGEEMG, rand.eff.b)
```

```
#N is the number of clusters
N<-length(countss$ID)
#p is the number of parameters
p<-length(betainGEEMG)
accuracy <- 0.001
error <-1
#itera is the number of iterations
itera <-0
while (error>accuracy){
I0 <-0; I1 <-0; I2 <-0
itera <-itera +1
for(i in 1:length(idua)){
```

#ni is the number of members in the ith cluster ni<-length(cluster[[i]]\$ID)</pre>

```
#nii is the number of elements in the ith column of matix D
nii < -(ni*(ni-1))/2+1
```

```
x.full <- cbind(1, cluster[[i]]$r, cluster[[i]]$x, cluster[[i]]$x*cluster[[i]]$r)
z.full <- cbind(1, cluster[[i]]$x)
y<-as.vector(cluster[[i]]$y)</pre>
```

```
q<-length(z.full[1,])
ul<- c((cluster[[i]]$y1)^2)
u2<-c((cluster[[i]]$y1[j])*(cluster[[i]]$y1[j+1]),
( cluster[[i]]$y1[j])*(cluster[[i]]$y1[j+2]),
(cluster[[i]]$y1[j])*(cluster[[i]]$y1[j+3]),
(cluster[[i]]$y1[j+1])*(cluster[[i]]$y1[j+2]),
(cluster[[i]]$y1[j+1])*(cluster[[i]]$y1[j+2]),
(cluster[[i]]$y1[j+2])*(cluster[[i]]$y1[j+2]),
```

```
u < -c(u1, u2)
```

```
mu < -list()
```

```
for (j in 1:nrow(x.full)){
mu[[j]] <- as.vector(x.full[j, ]%*%betainGEEMG +
t(z.full[j,])%*%G%*%z.full[j, ])
}</pre>
```

M1 <- diag(mu)

mu <- c(mu[[1]], mu[[2]], mu[[3]], mu[[4]])

```
lambda.i <- c(mu[[1]]^2, mu[[2]]^2, mu[[3]]^2, mu[[4]]^2, mu[[1]]*mu[[2]],
mu[[2]]*mu[[3]], mu[[3]]*mu[[4]])
M2 <- diag(lambda.i)</pre>
```


Update beta

zz.full <--matrix (c(z.full[1,]², z.full[2,]², z.full[3,]², z.full[4,]², z.full[1,]*z.full[2,], z.full[2,]*z.full[3,],z.full[3,]*z.full[4,]), 7, 2) DX<-M1%*%x.full DZ<-M2%*%zz.full

n . X < -p; n . Z < -q

```
 \begin{array}{l} D11 <- \operatorname{matrix}\left(\operatorname{rep}\left(0\,,\operatorname{ni}*\operatorname{n.X}\right), & \operatorname{nrow}=\operatorname{ni}\,, & \operatorname{ncol}=\operatorname{n.X}\right) \\ D12 <- \operatorname{matrix}\left(\operatorname{rep}\left(0\,,\operatorname{ni}*\operatorname{n.Z}\right), & \operatorname{nrow}=\operatorname{ni}\,, & \operatorname{ncol}=\operatorname{n.Z}\right) \\ D21 <- \operatorname{matrix}\left(\operatorname{rep}\left(0\,,\operatorname{nii}*\operatorname{n.X}\right), & \operatorname{nrow}=\operatorname{nii}\,, & \operatorname{ncol}=\operatorname{n.X}\right) \\ D22 <- \operatorname{matrix}\left(\operatorname{rep}\left(0\,,\operatorname{nii}*\operatorname{n.Z}\right), & \operatorname{nrow}=\operatorname{nii}\,, & \operatorname{ncol}=\operatorname{n.Z}\right) \\ D <- \operatorname{matrix}\left(\operatorname{rep}\left(0\,,(\operatorname{ni}+\operatorname{nii}\right)*\left(\operatorname{n.X+n.Z}\right)\right), & \operatorname{nrow}=\operatorname{ni}+\operatorname{nii}\,, & \operatorname{ncol}=\operatorname{n.X+n.Z}\right) \end{array}
```

```
V11<--matrix (rep(0,ni*ni), nrow=ni, ncol=ni)
V12<--matrix (rep(0,ni*nii), nrow=ni, ncol=nii)
V21<--matrix (rep(0,nii*ni), nrow=nii, ncol=ni)
V22<--matrix (rep(0,nii*nii), nrow=nii, ncol=nii)
V<--matrix (rep(0,(ni+nii)*(ni+nii)), nrow=ni+nii, ncol=ni+nii)</pre>
```

```
S11 \leftarrow matrix (rep (0, ni *1), nrow=ni, ncol=1)

S21 \leftarrow matrix (rep (0, nii *1), nrow=nii, ncol=1)

S \leftarrow matrix (rep (0, (ni+nii)*(1)), nrow=ni+nii, ncol=1)
```

```
 \begin{array}{l} H0 <- \mbox{ matrix} (\mbox{ rep}(0,(n.Z+n.X))*((n.Z+n.X))), \mbox{ nrow}=(n.Z+n.X), \mbox{ ncol}=(n.Z+n.X)) \\ H1 <- \mbox{ matrix} (\mbox{ rep}(0,(n.Z+n.X)), \mbox{ nrow}=(n.Z+n.X), \mbox{ ncol}=1) \end{array}
```

```
Ri.sigma2.delta <- matrix(rep(0, nii^2), nrow=nii)
diag(Ri.sigma2.delta)<-1
```

Ri. Matrix <-- matrix (0, nrow=ni, ncol=ni) diag(Ri. Matrix)<-1

T.<-4 J <- matrix (c(1), T., T.) Vi. Matrix <-M1%*%exp(z.ful1%*%G%*%t(z.ful1)-J)%*%M1+(M1) %*%Ri. Matrix

D11<-DX D12<-matrix (rep(0,ni*n.Z), nrow=ni, ncol=n.Z) D21<-matrix (rep(0,nii*n.X), nrow=nii, ncol=n.X) D22<-DZ D<-matrix (rep(0,(ni+nii)*(n.X+n.Z)), nrow=ni+nii, ncol=n.X+n.Z)

V11<-Vi. Matrix

```
V12<-matrix(rep(0,ni*nii), nrow=ni, ncol=nii)
V21<-matrix(rep(0,nii*ni), nrow=nii, ncol=ni)
V22<-Ri.sigma2.delta
V<-matrix(rep(0,(ni+nii)*(ni+nii)), nrow=ni+nii, ncol=ni+nii)</pre>
```

```
S11<-y-mu
S21<-u-lambda.i
S<-matrix (rep (0, (ni+nii)*(1)), nrow=ni+nii, ncol=1)
```

```
\begin{split} D[1:ni, 1:n.X] &< -D[1:ni, 1:n.X] + D11 \\ D[1:ni, (n.X+1):(n.X+n.Z)] &< -D[1:ni, (n.X+1):(n.X+n.Z)] + D12 \\ D[(1+ni):(ni+nii), 1:n.X] &< -D[(1+ni):(ni+nii), 1:n.X] + D21 \\ D[(1+ni):(ni+nii), (n.X+1):(n.X+n.Z)] &< -D[(1+ni):(ni+nii), (n.X+1):(n.X+n.Z)] \\ &< D[(1+ni):(n.X+n.Z)] + D22 \end{split}
```

```
V[1:ni, 1:ni]<-V[1:ni, 1:ni]+V11
V[1:ni, (ni+1):(ni+nii)]<-V[1:ni, (ni+1):(ni+nii)]+V12
V[(1+ni):(ni+nii), 1:ni]<-V[(1+ni):(ni+nii), 1:ni]+V21
V[(1+ni):(ni+nii), (ni+1):(ni+nii)]<-V[(1+ni):(ni+nii), (ni+1):(ni+nii)]+V22</pre>
```

```
S[1:ni, 1] < -S[1:ni, 1] + S11
S[(1+ni):(ni+nii), 1] < -S[(1+ni):(ni+nii), 1] + S21
```

```
H0<-H0+t (D)%*% ginv (V) %*%D
H1<-H1+t (D)%*% ginv (V) %*%(S)
```

```
10in <-- t (D)%*%V%*%D
```

```
I0 < -I0 + I0 in
Ilin <-- t (D)%*%V%*%S
I1 < -I1 + I1 in
I2in <-- t (D)%*%V%*%S%*%t (S)%*%V%*%D
I2 < -I2 + I2 in
}
U1<- U0 - ginv(H0) %*% H1
for (i \text{ in } 1: nrow(x.full))
mu.new <-as.vector(x.full[i, ]%*%beta.new + t(z.full[i,])%*%G%*%z.full[i, ])
}
\#rho <-((y [ i ]-mu.new)%*%t (y [ i ]-mu.new ))
rho <- 0
error < -sum((U1-U0)^2)
U0<-U1
mu <-mu.new
if (!(itera <25)) print ("Iterations > 25")
if (!(itera <25))
return(list(Converge="Error"))
}
```

```
beta00.GEEM2G[[1]]=U1[1]
beta11.GEEM2G[[1]]=U1[2]
beta22.GEEM2G[[1]]=U1[3]
beta33.GEEM2G[[1]]=U1[4]
```

```
var.b0[[1]]=U1[5]
var.b1[[1]]=U1[6]
}
betaGEEMG=c(mean(beta00.GEEM2G), mean(beta11.GEEM2G))
 mean(beta22.GEEM2G), mean(beta33.GEEM2G),
mean(var.b0), mean(var.b1))
sdGEEMG=c(sd(beta00.GEEM2G), sd(beta11.GEEM2G),
 sd(beta22.GEEM2G), sd(beta33.GEEM2G),
sd(var.b0), sd(var.b1))
seGEEMG=c(sdGEEMG[1]/sqrt(M.C), sqrt(sdGEEMG[2]/(M.C)),
sdGEEMG[3]/sqrt(M.C), sdGEEMG[4]/sqrt(M.C),
sdGEEMG[5]/sqrt(M.C), sdGEEMG[6]/sqrt(M.C))
betaFitGEEMG < -cbind(beta=betaGEEMG, S.D = (sdGEEMG))
 S = (seGEEMG)
colnames (betaFitGEEMG) <- c ("Estimate", "S.D", "S.E.")
Result=list (Beta=round (betaFitGEEMG, 10))
return (Result)
}
GEE. Mixed. Simu2 (sim. long. count1)
```