University of Louisville

ThinkIR: The University of Louisville's Institutional Repository

Electronic Theses and Dissertations

8-2008

Clustered longitudinal data analysis.

Ming Wang
University of Louisville

Follow this and additional works at: <https://ir.library.louisville.edu/etd>

Recommended Citation

Wang, Ming, "Clustered longitudinal data analysis." (2008). *Electronic Theses and Dissertations*. Paper 1524.
<https://doi.org/10.18297/etd/1524>

This Master's Thesis is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.

CLUSTERED LONGITUDINAL DATA ANALYSIS

By

Ming Wang

B.S., Mathematics, Peking University, CHINA, 2006

A Thesis

Submitted to the Faculty of the
Graduate School of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Master of Science

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

August 2008

CLUSTERED LONGITUDINAL DATA ANALYSIS

By

Ming Wang
B.S., Mathematics, Peking University, CHINA, 2006

A Thesis Approved on

06/19/2008

Date

By the following Thesis Committee:

Thesis Chair

Thesis Co-Chair

ACKNOWLEDGMENTS

I would like to thank my thesis advisors Dr. Maiying Kong and Dr. Somnath Datta for their insightful guidance, and Dr. Barbara A. Stetson for serving on my thesis committee. I also would like to thank Mr. Lijun Zhang for supporting me all the time. In addition, I would like to thank the Department of Bioinformatics and Biostatistics at University of Louisville for supporting me financially and academically, so that I can focus on my thesis research without extra difficulties in life. I appreciate all the faculty members and students in the department for their encouragement and friendship, particularly Dr. Guy Brock for helping me with R programming. I had a great time in my master's study during the past two years, and learned a lot from all the courses. The experience here has been the most valuable fortune for me and I will be always thankful for all those who helped me.

ABSTRACT

CLUSTERED LONGITUDINAL DATA ANALYSIS

Ming Wang

June 19, 2008

Clustered longitudinal data is often collected as repeated measurements on subjects over time arising in the clusters. Examples include longitudinal community intervention studies, or family studies with repeated measures on each member. Meanwhile, cluster size is sometime informative, which means that the risk for the outcomes is related to the cluster size. Under this situation, generalized estimating equations (GEE) will lead to invalid inferences because GEE assumes that the cluster size is non-informative.

In this study, we investigated the performances of generalized estimating equations (GEE), cluster-weighted generalized estimating equations (CWGEE), and within-cluster resampling (WCR) on clustered longitudinal data. Based on our extensive simulation studies, we conclude that all three methods provide comparable estimates when the cluster size is non-informative. But when cluster size is informative, GEE gives biased estimates, while WCR and CWGEE still provide unbiased and consistent estimates under different “working correlation structures” within-subject. However, WCR is a computationally intensive approach, so CWGEE is the best choice for clustered longitudinal data due to its solving only one estimating equation, which is asymptotically equivalent to WCR.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	iii
ABSTRACT	iv
LIST OF TABLES	viii
LIST OF FIGURES	x

CHAPTER

I	Introduction	1
	A Longitudinal data analysis	1
	B Clustered data analysis	3
	C Clustered longitudinal data analysis	4
	D Informative and noninformative cluster size	5
II	Clustered longitudinal data models	7
	A Generalized estimating equations	8
	B Within-cluster resampling	13
	C Cluster-weighted GEE model	14
	D Quasi-least squares method	16
III	Clustered longitudinal data with informative cluster size	19
	A Estimating parameters	19
	B Correlation structure	23
	1 AR-M: balanced	23
	2 Exchangeable: balanced	26

IV	Simulation studies	27
A	Simulation scenarios	27
1	Clustered longitudinal data with noninformative cluster size (balanced)	27
2	Clustered longitudinal data with informative cluster size I (balanced)	28
3	Clustered longitudinal data with informative cluster size II (balanced)	29
B	Simulation results and discussion	31
1	Tables (1-3) for clustered longitudinal data with noninfor- mative cluster size I ($N=50$)	31
2	Tables (4-6) for clustered longitudinal data with noninfor- mative cluster size II ($N=500$)	35
3	Tables (7-10) for clustered longitudinal data with informa- tive cluster size III (α_i , $N=50$)	38
4	Tables (11-14) for clustered longitudinal data with infor- mative cluster size IV (α_i and x_i , $N=50$)	42
C	Extension of clustered longitudinal data analysis	48
1	Tables (15-18) for clustered longitudinal data with infor- mative cluster size V (α_i , drop γ_{ij} , $N=50$)	48
2	Tables (19-22) for clustered longitudinal data with infor- mative cluster size VI (α_i and x_i , drop γ_{ij} , $N=50$)	48
3	Tables (23-24) for clustered longitudinal data with infor- mative cluster size VII (x_i , drop α_i and γ_{ij} , $N=50$)	48
D	Hypothesis test and power	56
V	Future work	59

REFERENCES	60
CURRICULUM VITAE	89

LIST OF TABLES

1	Different models for clustered longitudinal data with noninformative cluster size ($N=50$, $n=1000$ loops)	32
2	Variances and coverage rates for different models I ($N=50$, $n=1000$ loops)	33
3	Mean square error for different models I ($N=50$, $n=1000$ loops)	34
4	Mean square error for different models II ($N=500$, $n=1000$ loops)	35
5	Different models for clustered longitudinal data with noninformative cluster size ($N=500$, $n=1000$ loops)	36
6	Variances and coverage rates for different models II ($N=500$, $n=1000$ loops)	37
7	Different models for clustered longitudinal data with informative cluster size (α_i , $n=1000$ loops)	38
8	Variances and coverage rates for different models III (α_i , $n=1000$ loops)	39
9	Estimated correlation for informative clustered longitudinal data (α_i , $n=1000$ loops)	40
10	Mean square error for different models III (α_i , $n=1000$ loops)	40
11	Different models for clustered longitudinal data with informative cluster size (α_i and x_i , $n=1000$ loops)	43
12	Variances and coverage rates for different models IV (α_i and x_i , $n=1000$ loops)	44
13	Estimated correlation for clustered longitudinal data (α_i and x_i , $n=1000$ loops)	45

14	Mean square error for different models IV (α_i and x_i , $n=1000$ loops) .	45
15	Different models for clustered longitudinal data with informative cluster size (α_i and drop γ_{ij} , $n=1000$ loops)	49
16	Variances and coverage rates for different models V (α_i and drop γ_{ij} , $n=1000$ loops)	50
17	Estimated correlation for clustered longitudinal data with informative cluster size (α_i and drop γ_{ij} , $n=1000$ loops)	51
18	Mean square error for different models V (α_i and drop γ_{ij} , $n=1000$ loops)	51
19	Different models for clustered longitudinal data with informative cluster size (α_i and x_i , drop γ_{ij} , $n=1000$ loops)	52
20	Variances and coverage rates for different models VI (α_i and x_i , drop γ_{ij} , $n=1000$ loops)	53
21	Estimated correlation for clustered longitudinal data with informative cluster size (α_i and x_i , drop γ_{ij} , $n=1000$ loops)	54
22	Mean square error for different models VII (α_i and x_i , drop γ_{ij} , $n=1000$ loops)	54
23	Different models for clustered longitudinal data with informative cluster size (x_i , drop α_i and γ_{ij} , $n=1000$ loops)	55

LIST OF FIGURES

FIGURE	Page
TABLE	Page
1 Model relationship	2
2 Flow chart for clustered longitudinal data	7
3 Probability-Probability plots for different models I	41
4 Probability-Probability plots for CWGEE model I	42
5 Probability-Probability plots for different models II	46
6 Probability-Probability plots for CWGEE model II	46
7 Power plot for testing $H_0:\beta_2 = 0$ vs $H_a:\beta_2 \neq 0$	57
8 Power plot for testing $H_0:\beta_0 = 0$ vs $H_a:\beta_0 \neq 0$	58

CHAPTER I

Introduction

A Longitudinal data analysis

Longitudinal data analysis has gained increasing attentions in recent years [1]. The defining feature of longitudinal studies is that subjects are measured repeatedly over time. The observations from the same subject are usually correlated. For example, a child who is taller in the same age group tends to be taller one year later among the same age group. The techniques developed for longitudinal studies can be widely applied to panel studies, cohort studies, event history studies and time series analysis in various fields, such as society, epidemiology, and biology [1].

To our knowledge, if there is only one observation for each independent subject, statistical techniques such as general linear model or logistic regression model can be used. However, when the measurements are taken repeatedly on the same subject over time, the correlation within-subject should be taken into account to draw valid scientific inferences. So an extension of generalized linear models will be needed for longitudinal data analysis [2].

Over the past two decades, longitudinal data methods have been widely developed. However, under different situations, different statistical methods and assumptions are considered. For example, if the outcomes are continuous and normally distributed, mixed-effects linear model or generalized estimating equations (GEE) are commonly used [3]. If the outcomes are categorical, such as ordinal or nominal, more complicated nonlinear models and GEE may be used [4].

Furthermore, the simplest but more restrictive model is analysis of variance (ANOVA) for repeated measures, which assumes that the variance-covariance matrix is compound symmetric [5]. When multivariate analysis of variance (MANOVA) method is used, repeated measures are transformed into orthogonal polynomial coefficients which can be used as multivariate responses [6]. However, if the number of observations varies from one subject to another, more general methods, such as generalized mixed-effects regression models will be much more suitable [7].

Figure 1. Model relationship

In longitudinal studies, there are two types of covariates: non-time-varying and time-varying. Examples of non-time-varying covariates are gender, race and others, which remain constant over time; examples of time-varying covariates can be age, weight, income and so on, which may vary over time. If all subjects are measured at the same time, then this dataset can be referred to as equally-balanced data; if subjects are measured at different sets of times or there are missing data,

this dataset will be unequally-balanced data. In this thesis, equally-balanced data with non-time-varying covariates will be considered.

Longitudinal studies have many advantages. The primary one is that more efficient estimates can be obtained. In addition, fewer subjects are required to achieve a similar level of statistical power to cross-sectional studies, because more information is provided by repeated measures on each subject. The secondary advantage is that longitudinal data can provide information for individual change over time, which can be used to understand the heterogeneity of the population [7].

B Clustered data analysis

Clustered data is also referred to as multilevel data where data are collected in different clusters. The measurements from the same cluster are usually correlated because they share the same characteristics. For example, consider a study in which grip strength is measured on both hands of elderly twins at baseline and at one month post-baseline. Of course, this study is expected to yield data with three sources of correlation. As Kreft and De Leeuw point out [8], “the more individuals share common experiences due to closeness in space, the more they are similar, or to a certain extent, duplications of each other”. Longitudinal data can also be viewed as clustered data, where each subject can be referred to as a cluster.

In clustered analysis, the intra-cluster correlation coefficient (ICC) is used to measure the similarity within-cluster. Ignoring this correlation could lead to biased estimates, incorrect p-values and power. Although clustered data is often encountered in medical, biological and environmental studies, the statistical issues are still challenging. For example, sample size is inflated; how to calculate the effective sample size is not yet fully developed [9].

There exist many methods to analyze clustered data. Marginal, conditional,

and random-effects models are developed. Particularly, methods for continuous outcome with normal random errors are well developed due to the elegant properties of the normal distribution, which simplify model building and ease software development. However, categorical outcomes are prominent in statistical practices [10]. For example, quality of life outcomes are often scored on ordinal scales.

Based on literature about multilevel studies, we know that the pioneering work of Tett, Jackson and Rothstein [11] provided the consideration, development, and presentation of a random coefficient model, which is related to several traditional methods, such as variance components analysis [12]. However, these traditional models are of limited use because of restrictive assumptions concerning missing data across time and the variance-covariance structure of the repeated measures. To fully understand the change over time for specific individuals, hierarchical linear models (HLMs) [13] have been developed. Hierarchical generalized linear models (HGLMs) [14] were derived by extending HLMs. HGLMs provide a unified modeling framework to estimate cluster-specific quantities of interest, covariate effects, and components of variance [15]. In this situation, the normal theory can be applied by large-sample results, and also semi-parametric and parametric models can be widely referred [16].

C Clustered longitudinal data analysis

Clustered longitudinal data occurs when the measurements are taken on each subject over time, where subjects belong to different clusters. For example, in dental studies, repeated measurements are collected on each tooth from each patient over time. We can treat each patient as a cluster, and the health status for each tooth is measured over time. So two sources of correlations can be specified: observations from each tooth, and teeth from each patient. It is expected that the

measurements are more similar from the same tooth within the same patient. Another example is a longitudinal study where weight on siblings is taken over time. Many other examples of clustered longitudinal data are described in Goldstein [17]. It is noted that clustered longitudinal data can be applied to various areas including survival analysis, clinical trials, and spatial data analysis [6].

D Informative and noninformative cluster size

Noninformative cluster size means that the cluster size does not provide any information about the outcomes of interest. In another hand, informative cluster size means that the risk of outcomes is related to the cluster size [18]. In other words, the cluster size has effect on the distribution of the outcomes. This often arises in cluster-based design. In some epidemiologic genetic studies, families with more members having the disease are sampled. For example, it is common to sample families to study genetic susceptibility and its association with environmental factors. For some diseases, an individual with a positive family history is more susceptible to the disease. The Family Heart Study [19], which identifies genetic and non-genetic risk factors for coronary heart disease (CHD) is such an example [20]. Another example is the teeth study, where the number of teeth may be negatively related to disease status because people who are more susceptible to the disease may have lost more teeth. As a result, the cluster size (i.e., the number of teeth per subject) is informative.

GEE assumes that the cluster size is non-informative, and each individual observation contributes equally in the likelihood function. Thus, when the cluster size is informative, larger clusters are overweighted so that GEE leads to biased estimates. Recently, several models were generated for informative cluster size. Hoffman, Sen, and Weinberg [21] proposed a within-cluster resampling (WCR)

procedure, where extensive computation is required. Later on, Williamson, Datta, and Satten [22] proposed cluster-weighted generalized estimating equations (CWGEE) which modified GEE based on the WCR method, where the estimating equation is inversely weighted by cluster size. Follmann, Proschan, and Leifer [23] established the asymptotic theories and broad applications of the WCR method. Most recently, Benhin, Rao, and Scott [24] gave a comprehensive and deep discussion on the mean estimating equation approach on clustered data with informative cluster size. In this thesis, we will investigate the performance of GEE, WCR, and CWGEE models on clustered longitudinal data with non-informative and informative cluster size.

CHAPTER II

Clustered longitudinal data models

For clustered longitudinal data, without loss of generality, let Y_{ijk} represent the k^{th} response of j^{th} subject in i^{th} cluster, where $i = 1, 2, \dots, N$; $j=1,2,\dots,n_i$; $k=1,2,\dots,K_{ij}$. Associated with Y_{ijk} is X_{ijk} , which is a $p \times 1$ vector of covariates. Figure 2 illustrates clustered longitudinal data structure. In this thesis, we will do analysis based on simulated datasets.

Figure 2. Flow chart for clustered longitudinal data

In Chapter 1, we have given simple description of basic models for clustered or longitudinal data. In the past several years, WCR and CWGEE have been developed for clustered data with informative cluster size. These methods can be accomplished using R packages: glm and gee. In this thesis, we will adopt and extend these two methods to analyze clustered longitudinal data. GEE is also

applied and compared with WCR and CWGEE. We will illustrate the advantageous performance of WCR and CWGEE by performing extensive simulations.

A Generalized estimating equations

GEE is a general method of fitting statistical models for longitudinal data or clustered data based on quasi-likelihood function [25]. Quasi-likelihood method is an extension of maximum likelihood method, which is introduced by Wedderburn [26] and extended by McCullagh and Nelder [27]. GEE offers many advantages, not only suitable for various types of outcomes, but also flexible in incorporating different correlation structures [28].

In longitudinal study, let Y_{it} be a dependent variable for subject i at time t , and X_{it} is corresponding covariate, where $i = 1, 2, \dots, N$, and $t = 1, 2, \dots, n_i$. The observations for a subject or a cluster are correlated. We may write the data structure as follows:

Subject / Cluster:

$$\begin{array}{ccccccc}
 1 & (X_{11}, Y_{11}), & (X_{12}, Y_{12}), & \cdots & (X_{1n_1}, Y_{1n_1}) \\
 2 & (X_{21}, Y_{21}), & (X_{22}, Y_{22}), & \cdots & (X_{2n_2}, Y_{2n_2}) \\
 \vdots & \vdots & \vdots & \vdots & \vdots \\
 i & (X_{i1}, Y_{i1}), & (X_{i2}, Y_{i2}), & \cdots & (X_{in_i}, Y_{in_i}) \\
 \vdots & \vdots & \vdots & \vdots & \vdots \\
 N & (X_{N1}, Y_{N1}), & (X_{N2}, Y_{N2}), & \cdots & (X_{Nn_N}, Y_{Nn_N}).
 \end{array}$$

With the above notation, the responses in each row are correlated. Let us denote:

$$Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix}, \quad X_i = \begin{pmatrix} X'_{i1} \\ \vdots \\ X'_{in_i} \end{pmatrix} = \begin{pmatrix} X_{i,11} & \cdots & X_{i,1p} \\ \vdots & \vdots & \vdots \\ X_{i,n_i1} & \cdots & X_{i,n_ip} \end{pmatrix}.$$

The distribution of the response Y_{it} belongs to the exponential family. We get the following linear model:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_N \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_N \end{pmatrix} = X\beta + \varepsilon,$$

where β is a vector of the parameters. For a specific subject or cluster, we have the linear model as follows:

$$Y_i = X_i\beta + \varepsilon_i, \quad (i = 1, 2, \dots, N). \quad (1)$$

We denote $E(Y_i) = \mu_i$, and define a link function h which connects μ_i and $X_i\beta$:

$$h(\mu_i) = X_i\beta. \quad (2)$$

Similarly, the variance $Var(Y_i)$ is specified from a function g of μ_i , then we can write as follows:

$$Var(Y_i) = \frac{g(\mu_i)}{\phi}, \quad (3)$$

where ϕ is a scale parameter. In current study, we mainly consider the normal distribution. So the link function h is identity, and ϕ equals 1. Thus, Y_i is a multivariate normal random variable with mean vector $X_i\beta$ and variance-covariance matrix V_i which is $n_i \times n_i$ matrix. We know that the observations from different subjects or clusters are independent, then

$$Var(Y) = Var \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_N \end{pmatrix} = \begin{pmatrix} V_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & V_i & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & V_N \end{pmatrix} = BlockDiag(V_1, \cdots, V_i, \cdots, V_N). \quad (4)$$

The additional specification in GEE is the “working correlation structure” $R_i(\alpha)$, which is a $n_i \times n_i$ matrix for a given Y_i . If we assume the variance is homogeneous, and the correlation structure is assumed to be the same for clusters, then we will have:

$$Var(Y_i) = \begin{pmatrix} \sigma^2 & \sigma_{12} & \cdots & \cdots & \sigma_{1n_i} \\ \sigma_{12} & \sigma^2 & \sigma_{23} & \cdots & \sigma_{2n_i} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \sigma_{1n_i} & \sigma_{2n_i} & \cdots & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 R_i. \quad (5)$$

However, if the variance is not homogeneous, $var(\varepsilon_{ij}) = \sigma_{ij}^2$, then the variance for Y_i will be

$$Var(Y_i) = \begin{pmatrix} \sigma_{i1}^2 & \sigma_{12} & \cdots & \cdots & \sigma_{1n_i} \\ \sigma_{12} & \sigma_{i2}^2 & \sigma_{23} & \cdots & \sigma_{2n_i} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \sigma_{1n_i} & \sigma_{2n_i} & \cdots & \cdots & \sigma_{in_i}^2 \end{pmatrix} = A_i^{1/2} R_i(\alpha) A_i^{1/2}, \quad (6)$$

where

$$A_i = \begin{pmatrix} \sigma_{i1}^2 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{ij}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \sigma_{in_i}^2 \end{pmatrix} = \text{Diag}\{\sigma_{i1}^2, \dots, \sigma_{in_i}^2\}, \quad (7)$$

where A_i is $n_i \times n_i$ matrix with diagonal element of $\text{Var}(Y_{it})$ [3]. It is noted that the correlation matrix R_i is to describe the pattern of association of measurements within cluster. How to select R_i should be based on the best description of the scenario involved. This “working correlation structure” has many types, such as independence, exchangeable, AR-M, tri-diagonal and so on.

1. **Independence:** there is no correlation for the observations within each cluster, that is $R[k_1, k_2] = 0$, for $k_1 \neq k_2$, and R_i is a $n_i \times n_i$ identity matrix.
2. **Exchangeable:** exchangeable correlation structure assumes equality of correlations within each cluster, that is $R[k_1, k_2] = \alpha$ for $k_1 \neq k_2$; otherwise, 1. If R_i is $n_i \times n_i$ matrix, R_i will be positive definite for α in $(\frac{1}{n_i-1}, 1)$.
3. **AR-M:** AR-M correlation structure assumes that correlations will be smaller for measurements farther apart in terms of measurement occasion, that is $R[k_1, k_2] = \alpha^{|k_1 - k_2|}$ for α in $(-1, 1)$.
4. **Tri-diagonal correlation structure:** this structure assumes that $R[k_1, k_2] = \alpha$ for $|k_1 - k_2| = 1$, and $R[k_1, k_2] = 0$, otherwise.

In this thesis, we only consider the first three structures for clustered longitudinal data. For example, AR-M structure is plausible for longitudinal studies because this structure forces the correlation to decrease with increasing separation in measurements occasion.

In addition, the intuition of estimating the parameters through GEE is to “choose β so that $\mu(\beta)$ is close to Y_i on average and to optimally weight each residual $Y_i - \mu_i$ by the inverse of $Cov(Y_i)$ ” [29]. GEE provides $\widehat{\beta}_{GEE}$, which is asymptotic consistent even when the working correlation structure $R_i(\alpha)$ is misspecified. This indicates that GEE model can get asymptotically unbiased and consistent estimates of β even when the nature of intra-cluster correlation is unknown. Under mild regularity conditions, Liang and Zeger [2] asserted that $\sqrt{N}(\widehat{\beta}_{GEE} - \beta)$ is asymptotically normal with mean zero and “robust” variance-covariance matrix as follows:

$$Var(\widehat{\beta}_{GEE}) = N\theta_i^{-1}M_i\theta_i^{-1}, \quad (8)$$

where

$$\theta_i = \sum_{i=1}^N \left(\frac{\partial \mu_i}{\partial \beta} \right)' \widehat{Var}(Y_i)^{-1} \left(\frac{\partial \mu_i}{\partial \beta} \right), \quad (9)$$

$$M_i = \sum_{i=1}^N \left(\frac{\partial \mu_i}{\partial \beta} \right)' \widehat{Var}(Y_i)^{-1} S_i \widehat{Var}(Y_i)^{-1} \left(\frac{\partial \mu_i}{\partial \beta} \right), \quad (10)$$

$$S_i = (Y_i - \widehat{\mu}_i)(Y_i - \widehat{\mu}_i)'. \quad (11)$$

$\widehat{\beta}_{GEE}$ is consistent even when the correlation structure is misspecified (8). However, an incorrect specification of $R_i(\alpha)$ may affect the efficiency of the estimates through the size of $\widehat{Var}(Y_i)$. So if $R_i(\alpha)$ is the true one, then robust variance $\widehat{\beta}$ will be the same as naive variance $N\theta_i^{-1}$, while if it is not, then these two variances will not equal. Sometime, this can be helpful for us to select the correlation structure. Moreover, GEE has been implemented by many available softwares such as SAS, SPlus, and R.

We applied GEE to fit clustered longitudinal data with non-informative and informative cluster size with considering the correlation among the observations

within subject. It is well known that GEE is robust for misspecification of correlation structures, thus, all the subjects could be considered to be independent. When we apply GEE to clustered longitudinal data analysis, the estimating equation is:

$$U(\beta) = \sum_{i=1}^N \sum_{j=1}^{n_i} D'_{ij} V_{ij}^{-1} (Y_{ij} - u_{ij}) = 0, \quad (12)$$

where D_{ij} equals X_{ij} for identity link function and V_{ij} can be written as $A_{ij}^{\frac{1}{2}} R_{ij} A_{ij}^{\frac{1}{2}}$.

B Within-cluster resampling

Within-cluster resampling is a new method proposed by Hoffman to analyze clustered data [21]. This method is simple but computationally intensive. The asymptotic theory for WCR is general for various types of responses including continuous and discrete. Its advantage over GEE model is that it remains valid even when the cluster size is informative.

The notation for clustered data remains the same as before. WCR method randomly samples one observation from each of N subjects or clusters with replacement. Since N observations are independent, generalized linear model can be applied. This procedure can be repeated Q times, where Q is a large number. Thus, WCR estimator is achieved by averaging of the Q estimates. Let $\hat{\beta}_q$ denote the estimate for the q^{th} sampled dataset, $q = 1, 2, \dots, Q$, then the WCR estimator $\hat{\beta}_{wcr}$ can be written as:

$$\hat{\beta}_{wcr} = Q^{-1} \sum_{q=1}^Q \hat{\beta}_q. \quad (13)$$

WCR estimator is asymptotically normal, that is, as $N \rightarrow \infty$ and $Q \rightarrow \infty$, $N^{1/2}(\hat{\beta}_{wcr} - \beta) \sim N_p(0, V)$, where V is a finite and positive-definite matrix. This method is valid for analyzing data with informative cluster size due to the sampling

scheme. It is noted that WCR is cluster-based, where larger clusters are given the same weight as smaller ones because each resampling-based analysis uses a single observation to represent each cluster. Thus, the effects of informative cluster sizes are eliminated in WCR, and the marginal parameter will have a cluster-based interpretation. However, GEE gives more weight for larger clusters than smaller ones, and the difference in relative weighting does affect the asymptotic parameter when cluster size is informative.

In current work, we apply WCR to clustered longitudinal studies to examine whether this method still works. We randomly draw one subject from each of the N clusters with replacement to form Q samples. In each sample, the subjects are independent, and observations within each subject are correlated. Apparently, each sample is longitudinal dataset, then we can use GEE method to get a consistent estimates of the parameters. The estimating equation for q^{th} estimator is

$$S_q(\beta_q, \alpha) = \sum_{i=1}^N \left(\frac{\partial \mu_{iq}}{\partial \beta} \right)' Var(Y_{iq})^{-1} (Y_{iq} - \mu_{iq}) = 0. \quad (14)$$

We can get the asymptotic variance-covariance estimate for $\sqrt{N}(\hat{\beta}_{wcr} - \beta)$:

$$\hat{V} = \widehat{Var}\{\sqrt{N}(\hat{\beta}_{wcr} - \beta)\} = \frac{N}{Q} \sum_{i=1}^Q \hat{V}_q - \frac{N}{Q} \sum_{i=1}^Q (\hat{\beta}_q - \hat{\beta}_{wcr})(\hat{\beta}_q - \hat{\beta}_{wcr})'. \quad (15)$$

For clustered longitudinal data, the advantage of WCR method is to avoid specifying the correlation among the subjects within cluster. We only need to consider the correlation for the observations within subject. This method remains valid no matter whether the cluster size is informative or not.

C Cluster-weighted GEE model

Cluster-weighted GEE model is an alternative approach to fit marginal models when cluster size is informative, which is proposed by M. Williamson, Datta,

and Satten [22] This method is asymptotically equivalent to WCR. The advantage for this method is that it performs as well as WCR without extensive calculation [24].

Recall WCR method, for each one of Q resampled datasets, we can get $\widehat{\beta}_q$ from the estimating equation $S_q(\beta_q, \alpha) = 0$. We will average Q estimating equations to get the mean estimating equation for the parameters instead of averaging the $\widehat{\beta}_q$'s. Then the weighted estimating equation can be simply reduced to:

$$U(\beta, \alpha) = \frac{1}{Q} \sum_{q=1}^Q S_q(\beta_q, \alpha) = 0, \quad (16)$$

where

$$S_q(\beta_q, \alpha) = \sum_{i=1}^N U_{iq}(\beta, \alpha) = 0. \quad (17)$$

Clearly, as $Q \rightarrow \infty$ and $N \rightarrow \infty$, this method will converge to the expected value from WCR, so (16) and (17) will be equivalent to:

$$U(\beta, \alpha) = \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} U_{ij}(\beta, \alpha) = 0, \quad (18)$$

where

$$U_{ij}(\beta, \alpha) = \left(\frac{\partial \mu_{ij}}{\partial \beta} \right)' \widehat{Var}(Y_{ij})^{-1} (Y_{ij} - \mu_{ij}). \quad (19)$$

In original method, U_{ij} is only a number based on generalized linear model (GLM). However, in this thesis, U_{ij} is a vector based on generalized estimating equations (GEE). Moreover, $\sqrt{N}(\widehat{\beta} - \beta)$ is asymptotic normal distributed with mean zero and variance-covariance Φ that can be consistently estimated as follows:

$$\widehat{\Phi} = \widehat{\varphi}^{-1} \widehat{\nu} \widehat{\varphi}^{-1}, \quad (20)$$

where

$$\hat{\varphi} = \frac{1}{N} \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\partial U_{ij}(\beta, \alpha)}{\partial \beta} \Big|_{\beta=\hat{\beta}} \quad (21)$$

and

$$\hat{\nu} = \frac{1}{N} \sum_{i=1}^N \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} U_{ij}(\hat{\beta}, \alpha) \right\} \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} U_{ij}(\hat{\beta}, \alpha) \right\}'. \quad (22)$$

CWGEE provides similar estimates as WCR without intensive computations.

In our current work, we adopt and extend CWGEE to clustered longitudinal data and examine its performance.

D Quasi-least squares method

Quasi-least squares (QLS) is a two-stage computational approach to estimate the regression parameters β and the correlation coefficient α within the framework of GEE. Stage one of QLS for balanced and equally spaced data [30] and unbalanced and unequally spaced data [31] alternates between updating estimates of β and α until convergence.

It is known that QLS has the same estimating equations as GEE to estimate β . Based on this method, we will derive the estimating equations for CWGEE. Recall the notation for clustered longitudinal data, Y_{ijk} represent the k^{th} response of j^{th} subject in i^{th} cluster, where $i = 1, 2, \dots, N$, $j=1, 2, \dots, n_i$, $k=1, 2, \dots, K_{ij}$. X_{ijk} is a $p \times 1$ vector of covariates. We denote that $E(Y_{ijk}) = \mu_{ijk}$, and a link function h as we did before for GEE:

$$h(\mu_{ijk}) = X_{ijk}\beta. \quad (23)$$

In addition, we consider the “working correlation structure” within subjects, $R_{ij}(\alpha)$, which is a $K_{ij} \times K_{ij}$ matrix for a given Y_{ij} . Then the variance for Y_{ij} can be written as:

$$Var(Y_{ij}) = A_{ij}^{\frac{1}{2}} R_{ij}(\alpha) A_{ij}^{\frac{1}{2}}, \quad (24)$$

where A_{ij} is $K_{ij} \times K_{ij}$ matrix with $Var(Y_{ijk})$ as the k^{th} diagonal element. We will consider exchangeable and AR-M structures for $R_{ij}(\alpha)$. Let us denote

$\mu'_{ij} = (\mu_{ij1}, \mu_{ij2}, \dots, \mu_{ijK_{ij}})$, then the Pearson's residuals can be written as:

$$z_{ij}(\beta) = A_{ij}^{-\frac{1}{2}}(Y_{ij} - \mu_{ij}). \quad (25)$$

The generalized sum of squares for error is defined as

$$Q(\alpha, \beta) = \sum_{i=1}^N \left(\frac{1}{n_i}\right) \sum_{j=1}^{n_i} z'_{ij}(\beta) R_{ij}^{-1}(\alpha) z_{ij}(\beta). \quad (26)$$

Then we estimate the parameters α and β by minimizing $Q(\alpha, \beta)$. Note that

$$\frac{\partial Q(\alpha, \beta)}{\partial \beta} = 0 \Rightarrow \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \left(\frac{\partial \mu_{ij}}{\partial \beta}\right)' A_{ij}^{-\frac{1}{2}} R_{ij}(\alpha)^{-1} A_{ij}^{-\frac{1}{2}} (Y_{ij} - \mu_{ij}) = 0, \quad (27)$$

$$\frac{\partial Q(\alpha, \beta)}{\partial \alpha} = 0 \Rightarrow \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} z'_{ij}(\beta) \frac{\partial R_{ij}^{-1}(\alpha)}{\partial \alpha} z_{ij}(\beta) = 0. \quad (28)$$

From the above equations, we can see that QLS use the same estimating equations as CWGEE (18) to estimate β . In addition, QLS solves an estimating equation (28) for α . However, the solution in (28) may not be consistent [32]. A consistent estimate for α can be obtained from the following estimating equations:

$$\sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \text{trace} \left\{ \frac{\partial R_{ij}^{-1}(\delta)}{\partial \delta} R_{ij}(\alpha) \right\} |_{\delta=\hat{\alpha}} = 0. \quad (29)$$

The final estimate of β will be obtained by solving the estimating equation (27) with α replaced by $\hat{\alpha}_{qls}$. For clustered longitudinal data with informative cluster size, we can add weight term into the framework of GEE easily. Of course, the asymptotic distribution of $\hat{\beta}_{qls}$ is the same as the asymptotic distribution of $\hat{\beta}_{cwgee}$.

As a result, tests and confidence intervals for the parameters with QLS can also be implemented from above.

The following algorithm is used to estimate the parameters:

1. Obtain a starting value of $\hat{\beta}$ by assuming $\alpha = 0$ from the estimating equation (27). It is noted that the model with $\alpha = 0$ is equivalent to linear regression, logistic regression, or Poisson regression model for outcomes that are continuous, binary, or counts, respectively.
2. Get the Pearson's residuals with the current estimate of β , where the k^{th} Pearson's residual for subject j in cluster i is given by

$$z_{ijk} = \frac{Y_{ijk} - \hat{\mu}_{ijk}}{h(\hat{\mu}_{ijk})}. \quad (30)$$

3. Obtain the estimate of α by solving the estimating equation (28). Here, the “working correlation structure” should be specified first. The details are given in the next chapter.
4. Obtain the estimate of β by solving the estimating equation (27) again using the current estimate of α included in the pre-specified correlation structure.
5. After the convergence in stage one, update the estimate of α in stage two using the estimating equation (29). The solution for α depends on different correlation structure. The method of bisection is applied to solve the stage two estimating equation.
6. Obtain the estimate of β by solving (27) evaluated at $\hat{\alpha}_{qls}$ which is the stage two estimate of α from (29).
7. Repeat Steps 5 and 6 until convergence.

CHAPTER III

Clustered longitudinal data with informative cluster size

A Estimating parameters

Let us consider the following model:

$$Y_{ijk} = (x_i)' \beta_1 + z_{ijk} \beta_2 + \alpha_i + \gamma_{ij} + \varepsilon_{ijk}, \quad (31)$$

where $i = 1, \dots, N$, with N indicating the number of clusters; $j=1, \dots, n_i$, with n_i indicating the number of subjects for i^{th} cluster, and $k=1, \dots, k_{ij}$, with k_{ij} indicating the number of observations for subject j within i^{th} cluster. Here, cluster size is informative. At this moment, let us assume k_{ij} are all equal to the same number k . Thus, the design matrix:

$$X = \begin{bmatrix} X_1 & Z_1 \\ X_2 & Z_2 \\ \dots & \dots \\ X_N & Z_N \end{bmatrix},$$

where X_i is $n_i k \times p$ matrix with the following form:

$$X_i = \begin{bmatrix} x'_i \\ x'_i \\ \dots \\ x'_i \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

Assuming the correlation structure is independent, we can obtain that the estimate of the parameters as follows:

$$\hat{\beta} = (X'X)^{-1}X'Y, \quad (32)$$

where

$$\begin{aligned} X'X &= \begin{bmatrix} X'_1 & X'_2 & \dots & X'_N \\ Z'_1 & Z'_2 & \dots & Z'_N \end{bmatrix} \begin{bmatrix} X_1 & Z_1 \\ X_2 & Z_2 \\ \dots & \dots \\ X_N & Z_N \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^N n_i k x_i x'_i & \sum_{i=1}^N X_i Z_i \\ \sum_{i=1}^N Z'_i X_i & \sum_{i=1}^N Z'_i Z_i \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^N n_i k x_i x'_i & \sum_{i=1}^N X'_i Z_i \\ \sum_{i=1}^N Z'_i X_i & \sum_{i=1}^N Z'_i Z_i \end{bmatrix}. \end{aligned}$$

Thus, we can obtain the expected value for the estimates of the parameters $\hat{\beta}_l$ under the assumption that all observations are independent:

$$E(\hat{\beta}) = \beta + E \left\{ \begin{bmatrix} \sum_{i=1}^N n_i k x_i x'_i & \sum_{i=1}^N X'_i Z_i \\ \sum_{i=1}^N Z'_i X_i & \sum_{i=1}^N Z'_i Z_i \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^N n_i \alpha_i k x_i \\ \sum_{i=1}^N \alpha_i Z'_i J_{n_i k} \end{bmatrix} \right\}. \quad (33)$$

When cluster size is only related to cluster effect, Benhin, Rao and Scott [33] have proved that only the intercept is biased for the following model:

$$y_{ij} = x_i' \beta + \alpha_i + \varepsilon_{ij} \quad (j = 1, 2, \dots, n_i; \quad i = 1, 2, \dots, N). \quad (34)$$

In this model, x_i is a $p \times 1$ vector of fixed cluster specific covariate. When the cluster size is non-informative, $E(y_{ij}|x_i) = \mu_i = x_i' \beta$. However, when the cluster size is informative, $E(y_{ij}|x_i) = \mu_i = x_i' \beta + \alpha_i$, where α_i is related to the cluster size, and then the expected value will converge to:

$$\tilde{\beta} = \beta + \frac{E(\alpha h(\alpha))}{E(h(\alpha))} \left(\sum_{i=1}^k x_i x_i' \right)^{-1} \sum_{i=1}^k x_i. \quad (35)$$

By similar argument, the second term in (35) is not a zero vector. Particularly, the intercept term β_0 is biased. The simulations results in next chapter illustrate this point clearly.

In another case that cluster size n_i is not only related to cluster effect α_i , but also the cluster-specific covariate x_i , i.e., exposure factor in our model, we have $n_i = h(\alpha_i, x_i)$. Again, let assume all the observations are independent. The estimating equations estimator of β for the marginal model is given by:

$$\sum_{i=1}^N \sum_{j=1}^{n_i} \sum_{k=1}^k \begin{pmatrix} x_i \\ z_{ijk} \end{pmatrix} (Y_{ijk} - x_i' \beta_1 - z_{ijk}' \beta_2) = 0, \quad (36)$$

which implies that

$$\sum_{i=1}^N \sum_{j=1}^{n_i} \sum_{k=1}^k \begin{pmatrix} x_i \\ z_{ijk} \end{pmatrix} (x_i' \beta_1 + z_{ijk}' \beta_2) = \sum_{i=1}^N \sum_{j=1}^{n_i} \sum_{k=1}^k \begin{pmatrix} x_i \\ z_{ijk} \end{pmatrix} Y_{ijk}. \quad (37)$$

Thus,

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \left[\sum_{i=1}^N \sum_{j=1}^{n_i} \sum_{k=1}^k \begin{pmatrix} x_i \\ z_{ijk} \end{pmatrix} \begin{pmatrix} x_i' & z_{ijk}' \end{pmatrix} \right]^{-1} \sum_{i=1}^N \sum_{j=1}^{n_i} \sum_{k=1}^k \begin{pmatrix} x_i \\ z_{ijk} \end{pmatrix} Y_{ijk}, \quad (38)$$

and

$$\begin{aligned}
E \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} &= \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + E \left\{ \left[\sum_{i=1}^N \sum_{j=1}^{n_i} \sum_{k=1}^k \begin{pmatrix} x_i \\ z_{ijk} \end{pmatrix} (x'_i \ z'_{ijk}) \right]^{-1} \sum_{i=1}^N \sum_{j=1}^{n_i} \sum_{k=1}^k \begin{pmatrix} x_i \\ z_{ijk} \end{pmatrix} \alpha_i \right\} \\
&= \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + E \left\{ \left[\sum_{i=1}^N \begin{pmatrix} n_i k x_i x'_i & n_i k x_i \bar{z}'_{i..} \\ (n_i k \bar{z}_{i..}) x'_i & n_i k (zz')_{i..} \end{pmatrix} \right]^{-1} \sum_{i=1}^N \begin{pmatrix} n_i \alpha_i k x_i \\ \alpha_i (n_i k \bar{z}_{i..}) \end{pmatrix} \right\},
\end{aligned}$$

where

$$\bar{z}_{i..} = \frac{1}{n_i k} \sum_{j=1}^{n_i} \sum_{k=1}^k z_{ijk}, \quad (zz')_{i..} = \frac{1}{n_i k} \sum_{j=1}^{n_i} \sum_{k=1}^k z_{ijk} z'_{ijk}.$$

Since $n_i = h(\alpha_i, x_i)$, n_i does not depend on the subject-specific covariates. $\bar{z}_{i..}$ and $(zz')_{i..}$ are independent of n_i . Note that the second term in the expression may depend on the cluster size, thus it may not go to zero. Therefore some other coefficients of the covariates besides the intercept term β_0 may be also biased.

On the other hand, the estimating equation with independent correlation structure based on CWGEE can be written as:

$$\sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=1}^k \begin{pmatrix} x_i \\ z_{ijk} \end{pmatrix} (Y_{ijk} - x'_i \beta_1 - z'_{ijk} \beta_2) = 0. \quad (39)$$

That is

$$\sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=1}^k \begin{pmatrix} x_i \\ z_{ijk} \end{pmatrix} (x'_i \beta_1 + z'_{ijk} \beta_2) = \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=1}^k \begin{pmatrix} x_i \\ z_{ijk} \end{pmatrix} Y_{ijk}. \quad (40)$$

Then we get

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \left[\sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=1}^k \begin{pmatrix} x_i \\ z_{ijk} \end{pmatrix} (x'_i \ z'_{ijk}) \right]^{-1} \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=1}^k \begin{pmatrix} x_i \\ z_{ijk} \end{pmatrix} Y_{ijk}. \quad (41)$$

Thus, we can get the expected value:

$$\begin{aligned}
E \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} &= \beta + E \left\{ \left[\sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=1}^k \begin{pmatrix} x_i \\ z_{ijk} \end{pmatrix} \begin{pmatrix} x'_i & z'_{ijk} \end{pmatrix} \right]^{-1} \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=1}^k \begin{pmatrix} x_i \\ z_{ijk} \end{pmatrix} \alpha_i \right\} \\
&= \beta + E \left\{ \left[\sum_{i=1}^N \begin{pmatrix} kx_i x'_i & kx_i \bar{z}'_{i..} \\ (k\bar{z}_{i..})x'_i & k(zz')_{i..} \end{pmatrix} \right]^{-1} \sum_{i=1}^N \begin{pmatrix} \alpha_i kx_i \\ \alpha_i (k\bar{z}_{i..}) \end{pmatrix} \right\} \\
&= \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.
\end{aligned}$$

Since the second term does not depend on the cluster size any more and α_i is the only random variable in the expression with mean zero, the second term will be a zero vector. Thus the estimates of the parameters are unbiased in CWGEE.

B Correlation structure

In the simulation studies presented next chapter, we consider independent, exchangeable, and AR-M correlation structures. Quasi-least squares (QLS) method will be used to estimate α .

1 AR-M: balanced

We consider AR-M structure within each subject, i.e.

$$R_{ij}(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \dots & \alpha^{K-1} \\ \alpha & 1 & \alpha & \alpha^2 & \dots & \alpha^{K-2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \alpha^{K-1} & \alpha^{K-2} & \alpha^{K-3} & \alpha^{K-4} & \dots & 1 \end{pmatrix}.$$

Since R_{ij} is symmetric and positive definite matrix, we can use cholesky decomposition method to write R_{ij} as the product of a lower triangular matrix L_{ij}

and an upper triangular matrix L'_{ij} , i.e.

$$R_{ij}(\alpha) = L_{ij}(\alpha)L'_{ij}(\alpha), \quad (42)$$

where L'_{ij} is the transpose matrix of L_{ij} .

The formula for the entries of $L_{ij}(\alpha)$ is as follows:

$$L_{ij}(m, n) = \begin{cases} \alpha^{m-n}, & \text{if } m \geq n = 1; \\ \sqrt{1-\alpha^2}, & \text{if } m = n > 1; \\ \alpha^{m-n}\sqrt{1-\alpha^2}, & \text{if } m > n > 1; \\ 0, & \text{others,} \end{cases} \quad (43)$$

i.e.,

$$L_{ij}(\alpha) = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \alpha & \sqrt{1-\alpha^2} & 0 & \cdots & 0 \\ \alpha^2 & \alpha\sqrt{1-\alpha^2} & \sqrt{1-\alpha^2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha^{K-1} & \alpha^{K-2}\sqrt{1-\alpha^2} & \cdots & \cdots & \sqrt{1-\alpha^2} \end{pmatrix}. \quad (44)$$

It is clear that the inverse matrix of R_{ij} is also symmetric and positive definite.

Using the same method, we can get

$$R_{ij}^{-1}(\alpha) = T_{ij}(\alpha)T'_{ij}(\alpha), \quad (45)$$

where the entries of $T_{ij}(\alpha)$ have the following form:

$$T_{ij}(m, n) = \begin{cases} \frac{1}{\sqrt{1-\alpha^2}}, & \text{if } m = n < K; \\ \alpha^{m-n}, & \text{if } m = n = K; \\ \frac{-\alpha}{\sqrt{1-\alpha^2}}, & \text{if } m = n + 1; \\ 0, & \text{others.} \end{cases}$$

According to (45), we get

$$\frac{\partial R_{ij}^{-1}(\alpha)}{\partial \alpha} = \frac{\partial T_{ij}(\alpha) T'_{ij}(\alpha)}{\partial \alpha} = \frac{\partial T_{ij}(\alpha)}{\partial \alpha} T'_{ij}(\alpha) + T_{ij}(\alpha) \frac{\partial T'_{ij}(\alpha)}{\partial \alpha}. \quad (46)$$

Replace $\frac{\partial R_{ij}^{-1}(\alpha)}{\partial \alpha}$ in equation (28) by the expression on the right hand side of (46), we get the estimating equation for α :

$$\sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} (z_{ij1}, z_{ij2}, \dots, z_{ijK}) \frac{\partial R_{ij}^{-1}}{\partial \alpha} \begin{pmatrix} z_{ij1} \\ z_{ij2} \\ \dots \\ z_{ijK} \end{pmatrix} = 0, \quad (47)$$

where

$$\frac{\partial R_{ij}^{-1}}{\partial \alpha}(m, n) = \begin{cases} \frac{2\alpha}{(1-\alpha)^2}, & \text{if } m = n = 1, K; \\ \frac{4\alpha}{(1-\alpha)^2}, & \text{if } 1 < m = n < K; \\ \frac{-(1-\alpha)^2}{(1-\alpha)^2}, & \text{if } |m - n| = 1; \\ 0, & \text{others.} \end{cases}$$

Furthermore, a simple version of the estimating equations can be obtained by

$$t_1 \alpha^2 - t_2 \alpha + t_3 = 0, \quad (48)$$

where

$$t_1 = t_3 = \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=2}^K z_{ijk} z_{ijk-1}, \quad (49)$$

$$t_2 = \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=2}^K (z_{ijk}^2 + z_{ijk-1}^2). \quad (50)$$

Thus, we can get the estimate of α as follows:

$$\hat{\alpha} = \xi - \eta, \quad (51)$$

where

$$\xi = \frac{\sum_{i=1}^N \frac{1}{n_i} \sum_{i=1}^{n_i} \sum_{k=2}^K (z_{ijk}^2 + z_{ijk-1}^2)}{2 \sum_{i=1}^N \frac{1}{n_i} \sum_{i=1}^{n_i} \sum_{k=2}^K z_{ijk} z_{ijk-1}}, \quad (52)$$

$$\eta = \frac{\sqrt{\sum_{i=1}^N \frac{1}{n_i} \sum_{i=1}^{n_i} \sum_{k=2}^K (z_{ijk} + z_{ijk-1})^2 \sum_{i=1}^N \frac{1}{n_i} \sum_{i=1}^{n_i} \sum_{k=2}^K (z_{ijk} - z_{ijk-1})^2}}{2 \sum_{i=1}^N \frac{1}{n_i} \sum_{i=1}^{n_i} \sum_{k=2}^K z_{ijk} z_{ijk-1}}. \quad (53)$$

The stage two estimate of α in AR-M structure is given by

$$\hat{\alpha}_{qls} = \frac{2\hat{\alpha}}{1 + \hat{\alpha}^2}. \quad (54)$$

2 Exchangeable: balanced

Suppose the correlation matrix for observations within subject is exchangeable, then R_{ij} can be written as:

$$R_{ij}(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha & \cdots & \alpha \\ \alpha & 1 & \alpha & \cdots & \alpha \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha & \alpha & \alpha & \cdots & 1 \end{pmatrix}.$$

By using the same method as before, we can obtain the estimating equations for α :

$$\sum_{i=1}^N \frac{1}{n_i} \sum_{i=1}^{n_i} z'_{ij} z_{ij} - \sum_{i=1}^N \frac{1}{n_i} \sum_{i=1}^{n_i} \frac{1 + (K-1)\alpha^2}{(1 + (K-1)\alpha)^2} \left(\sum_{k=1}^K z_{ijk}(\beta) \right)^2 = 0. \quad (55)$$

The stage two estimate of α in the exchangeable structure is given by

$$\hat{\alpha}_{qls} = \frac{\hat{\alpha}(\hat{\alpha}(K-2) + 2)}{1 + \hat{\alpha}^2(K-1)}. \quad (56)$$

CHAPTER IV

Simulation studies

A Simulation scenarios

1 Clustered longitudinal data with noninformative cluster size (balanced)

Let Y_{ijk} denote the k^{th} response of j^{th} subject in i^{th} cluster. Associated with Y_{ijk} is X_{ijk} , which is a vector of covariates including time, predict factor and the interaction of these two variables. A mixed-effect model for Y_{ijk} is given by:

$$Y_{ijk} = (1, X_{ijk})\beta + \alpha_i + \gamma_{ij} + \varepsilon_{ijk}, \quad (57)$$

In simulation, we take mixed-effect model as the following form:

$$Y_{ijk} = \beta_0 + \beta_1 time + \beta_2 predict + \beta_3 time \times predict + \alpha_i + \gamma_{ij} + \varepsilon_{ijk}, \quad (58)$$

where $k = 1, 2, 3$; $j = 1, 2, \dots, n_i$; $i = 1, 2, \dots, N$ ($N = 50$ or 500);

α_i is random cluster effect, with $\alpha_i \sim N(0, 0.25)$;

γ_{ij} is random subject effect, with $\gamma_{ij} \sim N(0, 0.15)$;

ε_{ijk} is residual, with $\varepsilon_{ijk} \sim N(0, 0.01)$.

Noninformative cluster size can be generated as follows:

$$n_i = 1 + n_i^*, \quad \text{where } n_i^* \sim \text{Pois}(\exp(1.5)). \quad (59)$$

Clusters in this simulated data will be divided into two groups. The first $\frac{N}{2}$ clusters are exposed and the last $\frac{N}{2}$ clusters are unexposed. If the cluster is exposed, the predict variable for the subjects in this cluster will be assigned to 1; If the cluster is unexposed, the predict variable for each subject in this cluster will be assigned to 0. The underlying parameters $\beta = (0.5, 1.5, 0.75, 0.8)'$. The simulated clustered longitudinal data is balanced and complete, and also the cluster size is noninformative because cluster size is not related to the responses. We used GEE, WCR, and CWGEE models to fit this dataset, and then compare the results.

2 Clustered longitudinal data with informative cluster size I (balanced)

Let Y_{ijk} denote the k^{th} response of j^{th} subject in i^{th} cluster. Similarly, we can get the responses from mixed-effect model:

$$Y_{ijk} = \beta_0 + \beta_1 time + \beta_2 predict + \beta_3 time \times predict + \alpha_i + \gamma_{ij} + \varepsilon_{ijk}, \quad (60)$$

where $k = 1, 2, 3$; $j = 1, 2, \dots, n_i$; $i = 1, 2, \dots, N$ ($N = 50$);

α_i is random cluster effect, with $\alpha_i \sim N(0, 0.25)$;

γ_{ij} is random subject effect, with $\gamma_{ij} \sim N(0, 0.15)$;

ε_{ijk} is the residual, with $\varepsilon_{ijk} \sim N(0, 0.01)$.

The only difference from scenario I is that the cluster size is informative. Here, we consider two situations: (i) cluster size is related to α_i ; (ii) the cluster size is related to α_i and predict variable. We use the Poisson distribution to generate the cluster size.

1. Cluster size is related to α_i . Let us denote the parameter for the Poisson distribution as:

$$\lambda(\alpha_i) = \gamma_0 + \gamma_1 \alpha_i, \quad (61)$$

where $\gamma_0 = 3$ and $\gamma_1 = 5$. The cluster size is generated as:

$$n_i = 1 + n_i^*, \text{ where } n_i^* \sim \text{Pois}(\exp(\lambda(\alpha_i))). \quad (62)$$

2. Cluster size is related to α_i and predict variable x_i . Let us denote the parameter for the Poisson distribution is:

$$\lambda(\alpha_i, \text{predict}) = \gamma_0 + \gamma_1 \alpha_i - \gamma_2 \text{predict}, \quad (63)$$

where $\gamma_0 = 3$, $\gamma_1 = 5$, and $\gamma_2 = 5$. The i^{th} cluster size n_i is obtained by

$$n_i = 1 + n_i^*, \quad n_i^* \sim \text{Pois}(\exp(\lambda(\alpha_i, \text{predict}))). \quad (64)$$

3 Clustered longitudinal data with informative cluster size II (balanced)

We will change the underlying model to generate clustered longitudinal data with informative cluster size. Again, GEE, WCR, CWGEE models will be applied and compared their results.

1. We consider another mixed-effect model without cluster effect α_i and subject effect γ_{ij} as follows:

$$Y_{ijk} = \beta_0 + \beta_1 \text{time} + \beta_2 \text{predict} + \beta_3 \text{time} \times \text{predict} + \varepsilon_{ijk}, \quad (65)$$

where $k = 1, 2, 3$; $j = 1, 2 \dots n_i$; $i = 1, 2 \dots N$;

ε_{ijk} is the residual with multinormal distributed, $N(0, \sigma^2 \Sigma)$;

$\sigma^2 = 0.5$, $\Sigma = R_{ij}(0.9)$, where R_{ij} is assumed to be AR-M.

The cluster size in this model is only related to predict variable x_i , so n_i is generated as follows:

$$n_i = 1 + n_i^*, \quad n_i^* \sim \text{Pois}(\exp(\lambda(\text{predict}))), \quad (66)$$

where

$$\lambda(\text{predict}) = \tau \text{predict}, \quad \tau = -5. \quad (67)$$

2. We consider the mixed-effect model without subject effect γ_{ij} , and the response is given by:

$$Y_{ijk} = \beta_0 + \beta_1 \text{time} + \beta_2 \text{predict} + \beta_3 \text{time} \times \text{predict} + \alpha_i + \varepsilon_{ijk}, \quad (68)$$

where $k = 1, 2, 3; \quad j = 1, 2 \dots n_i; \quad i = 1, 2 \dots N;$

α_i is random cluster effect, with $\alpha_i \sim N(0, 0.25);$

ε_{ijk} is the residual with multinormal distribution, $N(0, \sigma^2 \Sigma);$

$\sigma^2 = 0.5, \Sigma = R_{ij}(0.9),$ where R_{ij} is taken as AR-M.

The cluster size is only related to cluster effect α_i :

$$n_i = 1 + n_i^*, n_i^* \sim \text{Pois}(\exp(\lambda(\alpha_i))), \quad (69)$$

where

$$\lambda(\alpha_i) = \gamma_0 + \gamma_1 \alpha_i, \quad (70)$$

with $\gamma_0 = 3$ and $\gamma_1 = 5.$

3. The model is the same as the previous one, but the cluster size is related to cluster effect α_i and predict factor, so n_i is generated based on:

$$n_i = 1 + n_i^*, \quad n_i^* \sim \text{Pois}(\exp(\lambda(\alpha_i, \text{predict}))), \quad (71)$$

where

$$\lambda(\alpha_i, \text{predict}) = \gamma_0 + \gamma_1 \alpha_i - \gamma_2 \text{predict}, \quad (72)$$

with $\gamma_0 = 3$, $\gamma_1 = 5$, and $\gamma_2 = 5$.

B Simulation results and discussion

1 Tables (1-3) for clustered longitudinal data with noninformative cluster size I (N=50)

From Table 1, we can see that for clustered longitudinal data with non-informative cluster size, GEE, WCR, and CWGEE all have unbiased estimates of the parameters, which can be obtained through comparing “*sd*” and “*rmse*”. Here, “*sd*” is standard deviation, i.e., standard error among estimates derived from simulations, and “*rmse*” is square root of mean square error. For each setting, we run n Monte Carlo simulations, which is 1000 in this thesis. That is,

$$sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{\theta}_i - \bar{\theta})^2}, \quad (73)$$

$$rmse = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\theta}_i - \theta_{true})^2}. \quad (74)$$

It is well known that, GEE still gives consistent estimates of β even when the “working correlation structure” is misspecified. Note that WCR, CWGEE are also based on the framework of GEE model, so under “independence”, “exchangeable”, and “AR-M” correlation structures, three models all get unbiased and consistent estimates, which can be verified by the results from Table 1.

TABLE 1

Different models for clustered longitudinal data with noninformative cluster size ($N=50$, $n=1000$ loops)

models	β_0	β_1	β_2	β_3	α
GEE (independence)	0.5018	1.5000	0.7480	0.8000	0
rmse	0.0543	0.0006	0.0782	0.0008	
sd	0.0542	0.0006	0.0782	0.0008	
GEE (exchangeable)	0.5018	1.5000	0.7480	0.8000	0.9987
rmse	0.0543	0.0006	0.0782	0.0008	
sd	0.0542	0.0006	0.0782	0.0008	
GEE (AR-M)	0.5018	1.5000	0.7480	0.8000	0.9987
rmse	0.0543	0.0006	0.0782	0.0008	
sd	0.0543	0.0006	0.0782	0.0008	
WCR (independence)	0.5021	1.5000	0.7477	0.8000	0
rmse	0.0515	0.0007	0.0707	0.0010	
sd	0.0515	0.0007	0.0707	0.0010	
WCR (exchangeable)	0.5021	1.5000	0.7479	0.8000	0.9987
rmse	0.0517	0.0007	0.0705	0.0010	
sd	0.0517	0.0007	0.0705	0.0010	0.0002
WCR (AR-M)	0.5021	1.5000	0.7479	0.8000	0.9987
rmse	0.0516	0.0007	0.0706	0.0010	
sd	0.0516	0.0007	0.0706	0.0010	0.0011
CWGEE (exchangeable)	0.5021	1.5000	0.7478	0.8000	0.9955
rmse	0.0514	0.0007	0.0704	0.0010	
sd	0.0514	0.0007	0.0704	0.0010	0.0221
CWGEE (AR-M)	0.5021	1.5000	0.7478	0.8000	0.9951
rmse	0.0514	0.0007	0.0704	0.0010	
sd	0.0514	0.0007	0.0704	0.0010	0.0346
True value	0.5	1.5	0.75	0.8	0.976

GEE: Generalized Estimating Equations; WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations; Empirical Standard Error, i.e, standard error among estimates derived from simulations: $sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{\theta}_i - \bar{\theta})^2}$; Square Root of Mean Square Error:

$$rmse = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\theta}_i - \theta_{true})^2};$$

TABLE 2

Variances and coverage rates for different models I ($N=50$, $n=1000$ loops)

Models	Clusters	Corr.str	parameters	robust.var	naive.var	coverage
GEE	50	Independent	b0	0.0012	0.0008	0.925
			b1	0.0005	0.0001	0.930
			b2	0.0017	0.0011	0.936
			b3	0.0008	0.0002	0.931
		Exchangeable	b0	0.0008	0.0007	0.925
			b1	0.0001	0.0001	0.930
			b2	0.0011	0.0012	0.936
			b3	0.0001	0.0001	0.931
		AR-M	b0	0.0008	0.0008	0.925
			b1	0.0001	0.0001	0.930
			b2	0.0011	0.0011	0.936
			b3	0.0001	0.0001	0.931
WCR	50	Independent	b0	0.0500		0.926
			b1	0.0007		0.934
			b2	0.0713		0.938
			b3	0.0009		0.929
		Exchangeable	b0	0.0500		0.926
			b1	0.0007		0.934
			b2	0.0713		0.938
			b3	0.0009		0.929
		AR-M	b0	0.0508		0.939
			b1	0.0007		0.931
			b2	0.0718		0.942
			b3	0.0009		0.927
CWGEE	50	Exchangeable	b0	0.0016	0.0067	0.942
			b1	0.0001	0.0013	0.941
			b2	0.0023	0.0094	0.951
			b3	0.0001	0.0018	0.942
		AR-M	b0	0.0016	0.0069	0.942
			b1	0.0001	0.0016	0.941
			b2	0.0023	0.0098	0.951
			b3	0.0001	0.0023	0.942

GEE: Generalized Estimating Equations; WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations;

TABLE 3

Mean square error for different models I ($N=50$, $n=1000$ loops)

Correlation Structure	GEE	WCR	CWGEE
independence	0.082 (0.01)	0.083 (0.01)	
exchangeable	0.082 (0.01)	0.083 (0.01)	0.082 (0.01)
AR-M	0.082 (0.01)	0.084 (0.01)	0.082 (0.01)

GEE: Generalized Estimating Equations; WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations; Mean square error: $mse = \frac{1}{n} \sum_{i=1}^n mse_i$; $mse_i = \frac{1}{n-4} \sum_{j=1}^N \sum_{k=1}^{n_i} (Y_{ijk} - \hat{Y}_{ijk})^2$

Also from Table 1-3, empirical standard errors for β' s from WCR and CWGEE models are close to each other, and are both less than those from GEE model. In addition, the coverage percentages for the parameters from CWGEE are much closer to 0.95 than those from GEE model. Still we can see that these three models have similar mean square error from Table 3, which is defined as below:

$$mse = \frac{1}{n} \sum_{i=1}^n mse_i, \quad (75)$$

with

$$mse_i = \frac{1}{n-4} \sum_{j=1}^N \sum_{k=1}^{n_i} (Y_{ijk} - \hat{Y}_{ijk})^2. \quad (76)$$

Using different models, we can also estimate the correlation coefficient. In our case, the underlying model has exchangeable correlation structure within subject.

The correlation coefficient α_{ij} for each subject should be as follows:

$$\alpha_{ij(k,k')} = \frac{Cov(Y_{ijk}, Y_{ijk'})}{\sqrt{Var(Y_{ijk})} \sqrt{Var(Y_{ijk'})}} = \frac{\sigma_\alpha^2 + \sigma_\gamma^2}{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\epsilon^2} = constant \quad (0.976). \quad (77)$$

From Table 1, all models get basically similar estimates of correlation coefficient, which imply high correlation for the observations within subject.

In summary, all three models provide unbiased estimates of parameters when the cluster size is noninformative. In terms of coverage rates, CWGEE provides more accurate coverage rates than WCR and GEE.

2 Tables (4-6) for clustered longitudinal data with noninformative cluster size II (N=500)

When the number of clusters, N , is larger, GEE, WCR, and CWGEE still provide unbiased estimates of the parameters. The difference among these models will be smaller comparing with the results from 50 clusters and those from 500 clusters. This can be also obtained by comparing “*rmse*” (74) with “*sd*” (73) from Table 4, or robust and naive variances of the parameters β' s from Table 5.

Apparently, the coverage percentages for the parameters from these three models are larger than those from scenario I, and they are close to the nominal coverage percentage 0.95.

In summary, GEE, WCR and CWGEE models have comparable performances as N increases. It is noted that these three models are equivalent to fit clustered longitudinal data with non-informative cluster size.

TABLE 4

Mean square error for different models II ($N=500$, $n=1000$ loops)

Correlation Structure	GEE	WCR	CWGEE
independence	0.085 (0.004)	0.085 (0.004)	
exchangeable	0.085 (0.004)	0.085 (0.004)	0.083 (0.009)
AR-M	0.085 (0.004)	0.085 (0.004)	0.083(0.009)

GEE: Generalized Estimating Equations; WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations; Mean Square: $mse = \frac{1}{n} \sum_{i=1}^n mse_i$; $mse_i = \frac{1}{n-4} \sum_{i=1}^N \sum_{j=1}^{n_i} \sum_{k=1}^k (Y_{ijk} - \hat{Y}_{ijk})^2$

TABLE 5

Different models for clustered longitudinal data with noninformative cluster size ($N=500$, $n=1000$ loops)

models	β_0	β_1	β_2	β_3	α
GEE (independence)	0.5018	1.5000	0.7479	0.8000	0
rmse	0.0175	0.0002	0.0244	0.0003	
sd	0.0174	0.0002	0.0243	0.0003	
GEE (exchangeable)	0.5018	1.5000	0.7479	0.8000	0.9988
rmse	0.0175	0.0002	0.0244	0.0003	
sd	0.0174	0.0002	0.0243	0.0003	0.0001
GEE (AR-M)	0.5018	1.5000	0.7479	0.8000	0.9988
rmse	0.0175	0.0002	0.0244	0.0003	
sd	0.0174	0.0002	0.0243	0.0003	0.0003
WCR (independence)	0.5005	1.5000	0.7507	0.7999	0
rmse	0.0134	0.0002	0.0197	0.0002	
sd	0.0134	0.0002	0.0197	0.0002	
WCR (exchangeable)	0.5005	1.5000	0.7507	0.7999	0.9987
rmse	0.0134	0.0002	0.0197	0.0002	
sd	0.0134	0.0002	0.0197	0.0002	0.0002
WCR (AR-M)	0.5005	1.5000	0.7507	0.7999	0.9988
rmse	0.0133	0.0002	0.0198	0.0002	
sd	0.0133	0.0002	0.0198	0.0002	0.0006
CWGEE (exchangeable)	0.5030	1.4999	0.7544	0.8000	0.9943
rmse	0.0113	0.0001	0.0185	0.0001	
sd	0.0113	0.0001	0.0185	0.0001	0.0034
CWGEE (AR-M)	0.5030	1.4999	0.7544	0.8000	0.9944
rmse	0.0115	0.0001	0.0185	0.0001	
sd	0.0114	0.0001	0.0185	0.0001	0.0036
True value	0.5	1.5	0.75	0.8	0.976

GEE: Generalized Estimating Equations; WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations; Empirical Standard Error, i.e. standard error among estimates derived from simulations: $sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{\theta}_i - \bar{\theta})^2}$; Square Root of Mean Square Error:

$$rmse = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\theta}_i - \theta_{true})^2};$$

TABLE 6

Variances and coverage rates for different models II ($N=500$, $n=1000$ loops)

Models	Clusters	Corr.str	parameters	robust.var	naive.var	coverage
GEE	500	Independent	b0	0.0004	0.0002	0.934
			b1	0.0003	0.0001	0.954
			b2	0.0005	0.0004	0.938
			b3	0.0003	0.0001	0.949
		Exchangeable	b0	0.0002	0.0002	0.934
			b1	0.0001	0.0001	0.954
			b2	0.0004	0.0004	0.938
			b3	0.0001	0.0001	0.949
		AR-M	b0	0.0002	0.0003	0.934
			b1	0.0001	0.0001	0.954
			b2	0.0004	0.0004	0.938
			b3	0.0001	0.0001	0.949
WCR	500	Independent	b0	0.0118		0.955
			b1	0.0002		0.926
			b2	0.0218		0.964
			b3	0.0003		0.929
		Exchangeable	b0	0.0118		0.955
			b1	0.0002		0.926
			b2	0.0218		0.964
			b3	0.0003		0.929
		AR-M	b0	0.0164		0.958
			b1	0.0002		0.915
			b2	0.0232		0.970
			b3	0.0003		0.924
CWGEE	500	Exchangeable	b0	0.0009	0.0015	0.973
			b1	0.0001	0.0009	0.959
			b2	0.0008	0.0016	0.967
			b3	0.0001	0.0008	0.953
		AR-M	b0	0.0009	0.0015	0.973
			b1	0.0001	0.0009	0.959
			b2	0.0008	0.0016	0.967
			b3	0.0001	0.0008	0.953

GEE: Generalized Estimating Equations; WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations;

3 Tables (7-10) for clustered longitudinal data with informative cluster size III (α_i , N=50)

TABLE 7

Different models for clustered longitudinal data with informative cluster size (α_i , $n=1000$ loops)

	β_0	β_1	β_2	β_3	α
GEE (independence)	0.7697	1.5000	0.7424	0.8000	0
rmse	0.2922	0.0002	0.1533	0.0003	
sd	0.1125	0.0002	0.1532	0.0003	
GEE (exchangeable)	0.7697	1.5000	0.7424	0.8000	0.9985
rmse	0.2922	0.0002	0.1533	0.0003	
sd	0.1125	0.0002	0.1532	0.0003	0.0003
GEE (AR-M)	0.7697	1.5000	0.7424	0.8000	0.9985
rmse	0.2922	0.0002	0.1533	0.0003	
sd	0.1125	0.0002	0.1532	0.0003	0.0004
WCR (independence)	0.4984	1.5000	0.7525	0.8000	0
rmse	0.0520	0.0005	0.0736	0.0007	
sd	0.0520	0.0005	0.0736	0.0007	
WCR (exchangeable)	0.4984	1.5000	0.7522	0.8000	0.9987
rmse	0.0519	0.0005	0.0736	0.0007	
sd	0.0519	0.0005	0.0736	0.0007	0.0002
WCR (AR-M)	0.4984	1.5000	0.7524	0.8000	0.9987
rmse	0.0519	0.0005	0.0735	0.0007	
sd	0.0519	0.0005	0.0735	0.0007	0.0009
CWGEE (exchangeable)	0.4991	1.5000	0.7515	0.8000	0.9973
rmse	0.0517	0.0004	0.0733	0.0006	
sd	0.0517	0.0004	0.0733	0.0006	0.0079
CWGEE (AR-M)	0.4992	1.5000	0.7515	0.8000	0.9973
rmse	0.0517	0.0004	0.0732	0.0006	
sd	0.0517	0.0004	0.0732	0.0006	0.0088
True value	0.5	1.5	0.75	0.8	0.9756

GEE: Generalized Estimating Equations; Empirical Standard Error, i.e., standard error among estimates derived from simulations: $sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{\theta}_i - \bar{\theta})^2}$; Square Root of Mean Square Error:

$$rmse = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\theta}_i - \theta_{true})^2};$$

For clustered longitudinal data with informative cluster size, when cluster

TABLE 8

Variances and coverage rates for different models III (α_i , $n=1000$ loops)

Models	Clusters	Corr.str	parameters	robust.var	naive.var	coverage
GEE	50	Independent	b0	0.0004	0.0003	0.003
			b1	0.0003	0.0001	0.945
			b2	0.0006	0.0004	0.928
			b3	0.0004	0.0001	0.948
		Exchangeable	b0	0.0003	0.0003	0.003
			b1	0.0001	0.0001	0.944
			b2	0.0004	0.0004	0.928
			b3	0.0001	0.0001	0.949
		AR-M	b0	0.0003	0.0003	0.002
			b1	0.0001	0.0001	0.951
			b2	0.0004	0.0004	0.929
			b3	0.0001	0.0001	0.959
WCR	50	Independent	b0	0.0486		0.920
			b1	0.0005		0.924
			b2	0.0687		0.938
			b3	0.0008		0.921
		Exchangeable	b0	0.0486		0.920
			b1	0.0005		0.924
			b2	0.0687		0.938
			b3	0.0008		0.921
		AR-M	b0	0.0492		0.930
			b1	0.0006		0.923
			b2	0.0638		0.932
			b3	0.0014		0.920
CWGEE	50	Exchangeable	b0	0.0016	0.0069	0.930
			b1	0.0001	0.0015	0.945
			b2	0.0022	0.0097	0.931
			b3	0.0001	0.0022	0.934
		AR-M	b0	0.0016	0.0066	0.930
			b1	0.0001	0.0012	0.945
			b2	0.0022	0.0094	0.932
			b3	0.0001	0.0017	0.934

GEE: Generalized Estimating Equations; WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations;

TABLE 9

Estimated correlation for informative clustered longitudinal data (α_i , $n=1000$ loops)

	α		α
WCR (exchangeable)	0.9987	CWGEE (exchangeable)	0.9253 * 0.9973 **
sd	0.0002	sd	0.0348 * 0.0079 **
WCR (AR-M)	0.9987	CWGEE (AR-M)	0.9376 * 0.9973 **
sd	0.0009	sd	0.0312 * 0.0088 **
True value	0.9756	True value	0.9756

WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations; * denotes the estimator $\hat{\alpha}$ of correlation coefficient on stage one; ** denotes the estimator $\hat{\alpha}_{qls}$ of correlation coefficient on stage two; Empirical Standard Error, i.e, standard error among estimates derived from simulations: $sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{\theta}_i - \bar{\theta})^2}$;

TABLE 10

Mean square error for different models III (α_i , $n=1000$ loops)

Correlation Structure	GEE	WCR	CWGEE
independence	0.069 (0.015)	0.083 (0.014)	
exchangeable	0.069 (0.015)	0.083 (0.014)	0.054 (0.012)
AR-M	0.069 (0.015)	0.083 (0.014)	0.054 (0.012)

GEE: Generalized Estimating Equations; WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations; Mean Square: $mse = \frac{1}{n} \sum_{i=1}^n mse_i$; $mse_i = \frac{1}{n-4} \sum_{i=1}^N \sum_{j=1}^{n_i} \sum_{k=1}^k (Y_{ijk} - \hat{Y}_{ijk})^2$

size is only related to cluster effect, only the intercept term β_0 in GEE model is biased from the simulation result, which agrees with the theoretical inference. The bias can be seen from the difference from “*rmse*” and “*sd*”, for the β' s. From Table 8, the “*rmse*” for the β_0 is 0.2922, while “*sd*” is 0.1125, and the coverage percentage is 0.003, indicating the estimate for β_0 is biased. In contrast, WCR and CWGEE models provide unbiased estimates and the coverage rates are close to nominal coverage rate 0.95. The “*sd*” and robust variances from CWGEE model are both

TABLE 2

Variances and coverage rates for different models I ($N=50$, $n=1000$ loops)

Models	Clusters	Corr.str	parameters	robust.var	naive.var	coverage
GEE	50	Independent	b0	0.0012	0.0008	0.925
			b1	0.0005	0.0001	0.930
			b2	0.0017	0.0011	0.936
			b3	0.0008	0.0002	0.931
		Exchangeable	b0	0.0008	0.0007	0.925
			b1	0.0001	0.0001	0.930
			b2	0.0011	0.0012	0.936
			b3	0.0001	0.0001	0.931
		AR-M	b0	0.0008	0.0008	0.925
			b1	0.0001	0.0001	0.930
			b2	0.0011	0.0011	0.936
			b3	0.0001	0.0001	0.931
WCR	50	Independent	b0	0.0500		0.926
			b1	0.0007		0.934
			b2	0.0713		0.938
			b3	0.0009		0.929
		Exchangeable	b0	0.0500		0.926
			b1	0.0007		0.934
			b2	0.0713		0.938
			b3	0.0009		0.929
		AR-M	b0	0.0508		0.939
			b1	0.0007		0.931
			b2	0.0718		0.942
			b3	0.0009		0.927
CWGEE	50	Exchangeable	b0	0.0016	0.0067	0.942
			b1	0.0001	0.0013	0.941
			b2	0.0023	0.0094	0.951
			b3	0.0001	0.0018	0.942
		AR-M	b0	0.0016	0.0069	0.942
			b1	0.0001	0.0016	0.941
			b2	0.0023	0.0098	0.951
			b3	0.0001	0.0023	0.942

GEE: Generalized Estimating Equations; WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations;

$$\frac{\partial R_{ij}^{-1}(\alpha)}{\partial \alpha} = \frac{\partial T_{ij}(\alpha) T'_{ij}(\alpha)}{\partial \alpha} = \frac{\partial T_{ij}(\alpha)}{\partial \alpha} T'_{ij}(\alpha) + T_{ij}(\alpha) \frac{\partial T'_{ij}(\alpha)}{\partial \alpha}. \quad (46)$$

Replace $\frac{\partial R_{ij}^{-1}(\alpha)}{\partial \alpha}$ in equation (28) by the expression on the right hand side of (46), we get the estimating equation for α :

$$\sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} (z_{ij1}, z_{ij2}, \dots, z_{ijK}) \frac{\partial R_{ij}^{-1}}{\partial \alpha} \begin{pmatrix} z_{ij1} \\ z_{ij2} \\ \dots \\ z_{ijK} \end{pmatrix} = 0, \quad (47)$$

where

$$\frac{\partial R_{ij}^{-1}}{\partial \alpha}(m, n) = \begin{cases} \frac{2\alpha}{(1-\alpha)^2}, & \text{if } m = n = 1, K; \\ \frac{4\alpha}{(1-\alpha)^2}, & \text{if } 1 < m = n < K; \\ \frac{-(1-\alpha)^2}{(1-\alpha)^2}, & \text{if } |m - n| = 1; \\ 0, & \text{others.} \end{cases}$$

Furthermore, a simple version of the estimating equations can be obtained by

$$t_1 \alpha^2 - t_2 \alpha + t_3 = 0, \quad (48)$$

where

$$t_1 = t_3 = \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=2}^K z_{ijk} z_{ijk-1}, \quad (49)$$

$$t_2 = \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=2}^K (z_{ijk}^2 + z_{ijk-1}^2). \quad (50)$$

Thus, we can get the estimate of α as follows:

$$\hat{\alpha} = \xi - \eta, \quad (51)$$

where

less than those from WCR model, thus, CWGEE model performs better than WCR model. This advantage can also be observed from mean square error of the outcome from Table 10, where CWGEE model has the smallest term.

Figure 3. Probability-Probability plots for different models I

In addition, we examined the goodness of fit of β_0 and β_2 under different models. Probability-probability (P-P) plots are used to examine whether the empirical coverage is close to the nominal coverage rate [34]. Here, we will vary α in $\text{seq}(0.01, 0.95, \text{by}=0.05)$, and then calculate the corresponding confidence interval $(1-\alpha)*100\%$, so we can get the percentage that the true value drop into each

confidence interval with specified α . From Figure 3, we can see that the plots of CWGEE and WCR models are close to the nominal level, while for GEE model, the plot of β_0 is far away from the nominal level. Figure 4 illustrates the P-P plot for CWGEE model under different correlation structures, indicating CWGEE model provides correct power calculation.

Figure 4. Probability-Probability plots for CWGEE model I

In summary, for clustered longitudinal data where cluster size is only related to the cluster effect, CWGEE and WCR perform better than GEE model in estimating the parameters. For the correlation coefficient, the estimate from CWGEE model is much closer to the true value, and also from mean square error for the responses, CWGEE is the least one, so CWGEE is a better choice than WCR.

4 Tables (11-14) for clustered longitudinal data with informative cluster size IV (α_i and x_i , N=50)

For clustered longitudinal data with informative cluster size, when cluster size is related to cluster effect α_i and exposed factor x_i , the intercept term β_0 and the parameter for the predict variable in GEE model are both biased based on the simulation results. From Table 11, the “*rmse*” for β_0 and β_2 are 0.2808 and 0.2403,

TABLE 11

Different models for clustered longitudinal data with informative cluster size (α_i and x_i , $n=1000$ loops)

	β_0	β_1	β_2	β_3	α
GEE (independence)	<u>0.7600</u>	<u>1.5000</u>	<u>0.5529</u>	<u>0.8000</u>	<u>0</u>
rmse	0.2808	0.0002	0.2403	0.0013	
sd	0.1063	0.0002	0.1375	0.0013	
GEE (exchangeable)	<u>0.7600</u>	<u>1.5000</u>	<u>0.5529</u>	<u>0.8000</u>	<u>0.9984</u>
rmse	0.2808	0.0002	0.2403	0.0013	
sd	0.1063	0.0002	0.1375	0.0013	0.0004
GEE (AR-M)	<u>0.7600</u>	<u>1.5000</u>	<u>0.5529</u>	<u>0.8000</u>	<u>0.9984</u>
rmse	0.2809	0.0002	0.2403	0.0013	
sd	0.1063	0.0002	0.1375	0.0013	0.0006
WCR (independence)	<u>0.4988</u>	<u>1.5000</u>	<u>0.7495</u>	<u>0.8000</u>	<u>0</u>
rmse	0.0519	0.0005	0.0785	0.0014	
sd	0.0519	0.0005	0.0785	0.0014	
WCR (exchangeable)	<u>0.4990</u>	<u>1.5000</u>	<u>0.7494</u>	<u>0.8000</u>	<u>0.9987</u>
rmse	0.0518	0.0005	0.0781	0.0014	
sd	0.0518	0.0005	0.0781	0.0014	0.0003
WCR (AR-M)	<u>0.5010</u>	<u>1.5010</u>	<u>0.7495</u>	<u>0.7972</u>	<u>0.9987</u>
rmse	0.0518	0.0006	0.0780	0.0014	
sd	0.0518	0.0006	0.0780	0.0014	0.0015
CWGEE (exchangeable)	<u>0.4984</u>	<u>1.5000</u>	<u>0.7522</u>	<u>0.8000</u>	<u>0.9628</u>
rmse	0.0517	0.0004	0.0733	0.0006	
sd	0.0517	0.0004	0.0733	0.0006	0.1524
CWGEE (AR-M)	<u>0.5003</u>	<u>1.4992</u>	<u>0.7476</u>	<u>0.8007</u>	<u>0.9622</u>
rmse	0.0518	0.0005	0.0738	0.0006	
sd	0.0518	0.0005	0.0738	0.0006	0.1634
True value	<u>0.5</u>	<u>1.5</u>	<u>0.75</u>	<u>0.8</u>	<u>0.9756</u>

GEE: Generalized Estimating Equations; Empirical Standard Error, i.e., standard error among estimates derived from simulations: $sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{\theta}_i - \bar{\theta})^2}$; Square Root of Mean Square Error:

$$rmse = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\theta}_i - \theta_{true})^2};$$

TABLE 12

Variances and coverage rates for different models IV (α_i and x_i , $n=1000$ loops)

Models	Clusters	Corr.str	parameters	robust.var	naive.var	coverage
GEE	50	Independent	b0	0.0004	0.0003	0.001
			b1	0.0003	0.0001	0.962
			b2	0.0023	0.0017	0.233
			b3	0.0010	0.0001	0.938
		Exchangeable	b0	0.0003	0.0003	0.001
			b1	0.0001	0.0001	0.962
			b2	0.0015	0.0017	0.233
			b3	0.0001	0.0001	0.938
		AR-M	b0	0.0003	0.0003	0.001
			b1	0.0001	0.0001	0.974
			b2	0.0015	0.0018	0.249
			b3	0.0001	0.0001	0.957
WCR	50	Independent	b0	0.0465		0.969
			b1	0.0014		0.983
			b2	0.0406		0.954
			b3	0.0020		0.989
		Exchangeable	b0	0.0465		0.971
			b1	0.0014		0.985
			b2	0.0405		0.955
			b3	0.0019		0.993
		AR-M	b0	0.0499		0.940
			b1	0.0005		0.914
			b2	0.0452		0.934
			b3	0.0014		0.921
CWGEE	50	Exchangeable	b0	0.0022	0.0080	0.940
			b1	0.0004	0.0030	0.941
			b2	0.0031	0.0113	0.940
			b3	0.0006	0.0042	0.952
		AR-M	b0	0.0022	0.0087	0.939
			b1	0.0004	0.0034	0.941
			b2	0.0031	0.0122	0.942
			b3	0.0006	0.0049	0.952

GEE: Generalized Estimating Equations; WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations;

TABLE 13

Estimated correlation for clustered longitudinal data (α_i and x_i , $n=1000$ loops)

	α		α
WCR (exchangeable)	0.9988	CWGEE (exchangeable)	0.8692 * 0.9628 **
sd	0.0003	sd	0.1827 * 0.1524 **
WCR (AR-M)	0.9987	CWGEE (AR-M)	0.8848 * 0.9622 **
sd	0.0015	sd	0.1786 * 0.1634 **
True value	0.9756	True value	0.9756

WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations; * denotes the estimator $\hat{\alpha}$ of correlation coefficient on stage one; ** denotes the estimator $\hat{\alpha}_{qls}$ of correlation coefficient on stage two; Empirical Standard Error, i.e., standard error among estimates derived from simulations: $sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{\theta}_i - \bar{\theta})^2}$;

TABLE 14

Mean square error for different models IV (α_i and x_i , $n=1000$ loops)

Correlation Structure	GEE	WCR	CWGEE
independence	0.068 (0.018)	0.084 (0.016)	
exchangeable	0.068 (0.018)	0.084 (0.016)	0.053 (0.021)
AR-M	0.068 (0.018)	0.084 (0.016)	0.053 (0.021)

GEE: Generalized Estimating Equations; WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations; Mean Square: $mse = \frac{1}{n} \sum_{i=1}^n mse_i$; $mse_i = \frac{1}{n-4} \sum_{i=1}^N \sum_{j=1}^{n_i} \sum_{k=1}^k (Y_{ijk} - \hat{Y}_{ijk})^2$

while the respective “ sd ” are 0.1063 and 0.1375. The “ $rmse$ ” and “ sd ” for the other parameters are almost the same. Based on (20), the coverage percentages for β_0 and β_2 are 0.001 and 0.233 (0.249 for AR-M), indicating the bias of the estimates. WCR and CWGEE models perform much better than GEE model in terms of providing unbiased estimates with the coverage rate close to the nominal coverage rate.

Similarly to the previous one, we can get the P-P plots for β_0 and β_2 under different models, Figures 5 and 6. The construction method is same as before, and

Figure 5. Probability-Probability plots for different models II

Figure 6. Probability-Probability plots for CWGEE model II

we notice that the plots of CWGEE and WCR models are close to the nominal level and have almost linear trend, while the plot of β_0 and β_2 for GEE model are far away from the nominal level.

In summary, for clustered longitudinal data with cluster size being related to the cluster effect and exposed factor, CWGEE and WCR models perform equally well to get unbiased estimates. While the “*sd*” and robust variances from CWGEE model are less than those from WCR model. In terms of the mean square error for the responses, CWGEE is the samllest. Thus, CWGEE has some advantage over WCR.

C Extension of clustered longitudinal data analysis

- 1 Tables (15-18) for clustered longitudinal data with informative cluster size V (α_i , drop γ_{ij} , N=50)
- 2 Tables (19-22) for clustered longitudinal data with informative cluster size VI (α_i and x_i , drop γ_{ij} , N=50)
- 3 Tables (23-24) for clustered longitudinal data with informative cluster size VII (x_i , drop α_i and γ_{ij} , N=50)

In this section, we will change the underlying model to extend the analysis of clustered longitudinal data with informative cluster size. From the estimates from different models, we can see that GEE model has biased estimates. WCR and CWGEE models provide unbiased estimates of the parameters by comparing the “*rmse*” with “*sd*”. The estimate of the correlation coefficient from WCR is 0.9066 when the “working correlation structure” is AR-M, which is close to the true value. The results are similar as before. Both WCR and CWGEE provide similar results for clustered longitudinal data with informative cluster size, however, WCR is computationally intensive.

TABLE 15

Different models for clustered longitudinal data with informative cluster size (α_i and drop γ_{ij} , $n=1000$ loops)

	β_0	β_1	β_2	β_3	α
GEE (independence)	<i>0.7751</i>	<i>1.4999</i>	<i>0.74227</i>	<i>0.7998</i>	<i>0</i>
rmse	0.2979	0.0068	0.1603	0.0094	
sd	0.1143	0.0068	0.1602	0.0094	
GEE (exchangeable)	<i>0.7751</i>	<i>1.4999</i>	<i>0.74227</i>	<i>0.7998</i>	<i>0.8809</i>
rmse	0.2979	0.0068	0.1603	0.0094	
sd	0.1143	0.0068	0.1602	0.0094	0.0054
GEE (AR-M)	<i>0.7752</i>	<i>1.4999</i>	<i>0.74227</i>	<i>0.7998</i>	<i>0.9084</i>
rmse	0.2980	0.0068	0.1604	0.0094	
sd	0.1143	0.0068	0.1603	0.0094	0.0048
WCR (independence)	<i>0.5010</i>	<i>1.5001</i>	<i>0.7459</i>	<i>0.8001</i>	<i>0</i>
rmse	0.0736	0.0148	0.1065	0.0204	
sd	0.0736	0.0148	0.1065	0.0204	
WCR (exchangeable)	<i>0.5010</i>	<i>1.5001</i>	<i>0.7459</i>	<i>0.8001</i>	<i>0.8799</i>
rmse	0.0732	0.0142	0.1081	0.0199	
sd	0.0733	0.0142	0.1081	0.0200	0.0107
WCR (AR-M)	<i>0.5005</i>	<i>1.4997</i>	<i>0.7472</i>	<i>0.8004</i>	<i>0.9067</i>
rmse	0.0717	0.0144	0.1054	0.0200	
sd	0.0717	0.0144	0.1054	0.0200	0.0098
CWGEE (exchangeable)	<i>0.5012</i>	<i>1.4998</i>	<i>0.7464</i>	<i>0.8004</i>	<i>0.8566</i>
rmse	0.0696	0.0134	0.1019	0.0185	
sd	0.0696	0.0134	0.1019	0.0185	0.1098
CWGEE (AR-M)	<i>0.5012</i>	<i>1.4998</i>	<i>0.7465</i>	<i>0.8004</i>	<i>0.8881</i>
rmse	0.0695	0.0134	0.1017	0.0185	
sd	0.0695	0.0134	0.1017	0.0185	0.0955
True value	<i>0.5</i>	<i>1.5</i>	<i>0.75</i>	<i>0.8</i>	<i>0.9</i>

GEE: Generalized Estimating Equations; Empirical Standard Error, i.e, standard error among estimates derived from simulations: $sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{\theta}_i - \bar{\theta})^2}$; Square Root of Mean Square Error:

$$rmse = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\theta}_i - \theta_{true})^2};$$

TABLE 16

Variances and coverage rates for different models V (α_i and drop γ_{ij} , $n=1000$ loops)

Models	Clusters	Corr.str	parameters	robust.var	naive.var	coverage
GEE	50	Independent	b0	0.0011	0.0008	0.018
			b1	0.0005	0.0002	0.941
			b2	0.0016	0.0012	0.915
			b3	0.0007	0.0003	0.956
		Exchangeable	b0	0.0008	0.0008	0.018
			b1	0.0002	0.0002	0.941
			b2	0.0011	0.0012	0.915
			b3	0.0003	0.0003	0.956
		AR-M	b0	0.0008	0.0008	0.009
			b1	0.0002	0.0002	0.954
			b2	0.0012	0.0012	0.915
			b3	0.0003	0.0003	0.957
WCR	50	Independent	b0	0.0783		0.928
			b1	0.0185		0.923
			b2	0.1105		0.925
			b3	0.0256		0.918
		Exchangeable	b0	0.0783		0.928
			b1	0.0185		0.923
			b2	0.1105		0.925
			b3	0.0256		0.918
		AR-M	b0	0.0781		0.925
			b1	0.0182		0.921
			b2	0.1080		0.927
			b3	0.0260		0.919
CWGEE	50	Exchangeable	b0	0.0022	0.0080	0.931
			b1	0.0004	0.0030	0.945
			b2	0.0031	0.0113	0.933
			b3	0.0006	0.0043	0.957
		AR-M	b0	0.0022	0.0087	0.936
			b1	0.0004	0.0035	0.945
			b2	0.0031	0.0123	0.934
			b3	0.0006	0.0049	0.957

GEE: Generalized Estimating Equations; WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations;

TABLE 17

Estimated correlation for clustered longitudinal data with informative cluster size (α_i and drop γ_{ij} , $n=1000$ loops)

	α		α
WCR (exchangeable)	0.8799	CWGEE (exchangeable)	0.5462 * 0.8566 **
sd	0.0107	sd	0.1009 * 0.1098 **
WCR (AR-M)	0.9067	CWGEE (AR-M)	0.6251 * 0.8880 **
sd	0.0098	sd	0.0946 * 0.0955 **
True value	0.9	True value	0.9

WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations; * denotes the estimator $\hat{\alpha}$ of correlation coefficient on stage one; ** denotes the estimator $\hat{\alpha}_{qls}$ of correlation coefficient on stage two; Empirical Standard Error, i.e, standard error among estimates derived from simulations: $sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{\theta}_i - \bar{\theta})^2}$;

TABLE 18

Mean square error for different models V (α_i and drop γ_{ij} , $n=1000$ loops)

Correlation Structure	GEE	WCR	CWGEE
independence	0.547 (0.022)	0.557 (0.043)	
exchangeable	0.547 (0.022)	0.557 (0.043)	0.558 (0.103)
AR-M	0.547 (0.022)	0.556 (0.040)	0.558 (0.103)

GEE: Generalized Estimating Equations; WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations; Mean Square: $mse = \frac{1}{n} \sum_{i=1}^n mse_i$; $mse_i = \frac{1}{n-4} \sum_{j=1}^N \sum_{k=1}^{n_i} (Y_{ijk} - \hat{Y}_{ijk})^2$

TABLE 19

Different models for clustered longitudinal data with informative cluster size (α_i and x_i , drop γ_{ij} , $n=1000$ loops)

	β_0	β_1	β_2	β_3	α
GEE (independence)	<u>0.7663</u>	<u>1.4999</u>	<u>0.5452</u>	<u>0.7999</u>	<u>0</u>
rmse	0.2888	0.0068	0.2844	0.0392	
sd	0.1119	0.0068	0.1975	0.0392	
GEE (exchangeable)	<u>0.7663</u>	<u>1.4999</u>	<u>0.5452</u>	<u>0.7999</u>	<u>0.8805</u>
rmse	0.2888	0.0068	0.2844	0.0392	
sd	0.1119	0.0068	0.1975	0.0392	0.0073
GEE (AR-M)	<u>0.7663</u>	<u>1.4999</u>	<u>0.5452</u>	<u>0.7999</u>	<u>0.9081</u>
rmse	0.2888	0.0068	0.2840	0.0392	
sd	0.1118	0.0068	0.1968	0.0392	0.0067
WCR (independence)	<u>0.5030</u>	<u>1.5002</u>	<u>0.7473</u>	<u>0.7991</u>	<u>0</u>
rmse	0.0732	0.0149	0.1793	0.0439	
sd	0.0732	0.0149	0.1793	0.0439	
WCR (exchangeable)	<u>0.5030</u>	<u>1.5002</u>	<u>0.7473</u>	<u>0.7991</u>	<u>0.879</u>
rmse	0.0732	0.0149	0.1793	0.0439	
sd	0.0732	0.0149	0.1793	0.0439	0.0206
WCR (AR-M)	<u>0.5030</u>	<u>1.5002</u>	<u>0.7473</u>	<u>0.7991</u>	<u>0.9054</u>
rmse	0.0744	0.0153	0.1798	0.0433	
sd	0.0744	0.0153	0.1798	0.0433	0.0193
CWGEE (exchangeable)	<u>0.4979</u>	<u>1.500</u>	<u>0.7529</u>	<u>0.7999</u>	<u>0.8553</u>
rmse	0.0238	0.0132	0.1677	0.0158	
sd	0.0238	0.0132	0.1677	0.0158	0.1891
CWGEE (AR-M)	<u>0.4981</u>	<u>1.4999</u>	<u>0.7529</u>	<u>0.7999</u>	<u>0.8581</u>
rmse	0.0238	0.0123	0.1676	0.0147	
sd	0.0238	0.0123	0.1676	0.0147	0.1512
True value	<u>0.5</u>	<u>1.5</u>	<u>0.75</u>	<u>0.8</u>	<u>0.9</u>

GEE: Generalized Estimating Equations; Empirical Standard Error, i.e, standard error among estimates derived from simulations: $sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{\theta}_i - \bar{\theta})^2}$; Square Root of Mean Square Error:

$$rmse = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\theta}_i - \theta_{true})^2};$$

TABLE 20

Variances and coverage rates for different models VI (α_i and x_i , drop γ_{ij} , $n=1000$ loops)

Models	Clusters	Corr.str	parameters	robust.var	naive.var	coverage
GEE	50	Independent	b0	0.0011	0.0008	0.016
			b1	0.0005	0.0002	0.954
			b2	0.0064	0.0047	0.676
			b3	0.0030	0.0012	0.934
		Exchangeable	b0	0.0008	0.0008	0.016
			b1	0.0002	0.0002	0.954
			b2	0.0045	0.0047	0.676
			b3	0.0010	0.0012	0.934
		AR-M	b0	0.0008	0.0008	0.011
			b1	0.0002	0.0002	0.947
			b2	0.0047	0.0040	0.586
			b3	0.0012	0.0011	0.927
WCR	50	Independent	b0	0.0785		0.924
			b1	0.01805		0.921
			b2	0.1749		0.926
			b3	0.0441		0.930
		Exchangeable	b0	0.0785		0.924
			b1	0.01805		0.921
			b2	0.1749		0.926
			b3	0.0441		0.930
		AR-M	b0	0.0781		0.924
			b1	0.0169		0.913
			b2	0.1765		0.927
			b3	0.0445		0.926
CWGEE	50	Exchangeable	b0	0.0016	0.0063	0.935
			b1	0.0001	0.0005	0.936
			b2	0.0022	0.0089	0.939
			b3	0.0001	0.0007	0.934
		AR-M	b0	0.0015	0.0067	0.935
			b1	0.0001	0.0009	0.943
			b2	0.0022	0.0095	0.940
			b3	0.0001	0.0013	0.934

GEE: Generalized Estimating Equations; WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations;

TABLE 21

Estimated correlation for clustered longitudinal data with informative cluster size (α_i and x_i , drop γ_{ij} , $n=1000$ loops)

	α		α
WCR (exchangeable)	0.8793	CWGEE (exchangeable)	0.5979 * 0.8553 **
sd	0.0206	sd	0.1179 * 0.1891 **
WCR (AR-M)	0.9054	CWGEE (AR-M)	0.5666 * 0.8581 **
sd	0.0193	sd	0.0917 * 0.1512 **
True value	0.9	True value	0.9

WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations; * denotes the estimator $\hat{\alpha}$ of correlation coefficient on stage one; ** denotes the estimator $\hat{\alpha}_{qls}$ of correlation coefficient on stage two; Empirical Standard Error, i.e, standard error among estimates derived from simulations: $sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{\theta}_i - \bar{\theta})^2}$;

TABLE 22

Mean square error for different models VII (α_i and x_i , drop γ_{ij} , $n=1000$ loops)

Correlation Structure	GEE	WCR	CWGEE
independence	0.546 (0.029)	0.553 (0.080)	
exchangeable	0.546 (0.029)	0.553 (0.080)	0.574 (0.039)
AR-M	0.546 (0.029)	0.553 (0.080)	0.574 (0.039)

GEE: Generalized Estimating Equations; WCR: Within-Cluster Resampling; CWGEE: Cluster-Weighted Generalized Estimating Equations; Mean Square: $mse = \frac{1}{n} \sum_{i=1}^n mse_i$; $mse_i = \frac{1}{n-4} \sum_{j=1}^N \sum_{k=1}^{n_i} (Y_{ijk} - \hat{Y}_{ijk})^2$

TABLE 23

Different models for clustered longitudinal data with informative cluster size (x_i , drop α_i and γ_{ij} , $n=1000$ loops)

	β_0	β_1	β_2	β_3
GEE (independence)	0.497	1.5	0.753	0.8
rmse	0.1135	0.0303	0.1141	0.0305
sd	0.1135	0.0303	0.1141	0.0305
GEE (exchangeable)	0.497	1.5	0.753	0.8
rmse	0.1135	0.0303	0.1141	0.0305
sd	0.1135	0.0303	0.1141	0.0305
GEE (AR-M)	0.497	1.5	0.753	0.8
rmse	0.1132	0.0303	0.1137	0.0305
sd	0.1132	0.0303	0.1137	0.0305
CWGEE (independence)	0.5	1.5	0.754	0.8
rmse	0.1339	0.0358	0.1340	0.0358
sd	0.1339	0.0358	0.1340	0.0358
CWGEE (exchangeable)	0.495	1.5	0.754	0.8
rmse	0.1337	0.0358	0.1339	0.0358
sd	0.1337	0.0358	0.1339	0.0358
CWGEE (AR-M)	0.495	1.5	0.754	0.8
rmse	0.1334	0.0358	0.1337	0.0358
sd	0.1334	0.0358	0.1337	0.0358
True value	0.5	1.5	0.75	0.8

GEE: Generalized Estimating Equations; CWGEE: Cluster-Weighted Generalized Estimating; Empirical Standard Error, i.e., standard error among estimates derived from simulations: $sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{\theta}_i - \bar{\theta})^2}$; Square Root of Mean Square Error: $rmse = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\theta}_i - \theta_{true})^2}$;

D Hypothesis test and power

From above simulations, for clustered longitudinal data, CWGEE performs as well as WCR but requires less computation, so we prefer CWGEE. We will do hypothesis test for the parameters based on CWGEE, and get the power for β_0 and β_2 using robust variance under the scenario where the cluster size is related to cluster effect α_i .

We choose the “working” correlation structure R_{ij} as exchangeable and AR-M respectively. The procedure is that under each type of R_{ij} , we will get the power for testing $H_0:\beta_2 = 0$ versus $H_a:\beta_2 \neq 0$ with the underlying value, δ varying from $\text{seq}(-1,1, \text{by}=0.05)$. To do that, we fix the other parameters as before. For each fixed β_2 , we draw 1000 samples from underlying model. For each sample, we use CWGEE model to obtain the estimators for all the parameters and their standard errors. Then we use Wald test to test $H_0:\beta_2 = 0$. We will calculate the percentage of the H_0 being rejected, which will be the power for $H_a:\beta_2 \neq 0$.

$$\text{power}^i(\delta) = \frac{\sum_{j=1}^{1000} I_{\text{reject}:H_0}}{1000}. \quad (78)$$

The results are illustrated in Figure 7. The procedures for β_0 are the same as β_2 , which is presented in Figure 8. Also we can get the confidence interval for power^i :

$$\left(\text{power}^i - 1.96 * \sqrt{\frac{\text{power}^i * (1 - \text{power}^i)}{1000}}, \quad \text{power}^i + 1.96 * \sqrt{\frac{\text{power}^i * (1 - \text{power}^i)}{1000}} \right). \quad (79)$$

Figure 7 showed that with the power for β_2 under exchangeable and AR-M correlation structures, the lowest power is 0.059, and the confidence interval is (0.044, 0.074) which includes 0.05; Figure 8 showed that with the power for β_0 under exchangeable and AR-M correlation structures, the lowest power is 0.03, and the confidence interval is (0.02, 0.04). So we can see that CWGEE model is an efficient method to estimate the parameters.

Figure 7. Power plot for testing $H_0: \beta_2 = 0$ vs $H_a: \beta_2 \neq 0$

power for AR-M and Exchangeable correlation structures

Figure 8. Power plot for testing $H_0:\beta_0 = 0$ vs $H_a:\beta_0 \neq 0$

CHAPTER V

Future work

CWGEE is an efficient method for clustered longitudinal data with informative cluster size. It provides similar estimates as WCR, which is an intensively computational method. In the future, we will investigate how to handle missing data in CWGEE model for clustered longitudinal studies. Note that there are different types of missing mechanism, such as missing not at random, missing at random and missing completely at random. Another is that each subject has the same number of observations in this thesis, but when the time points are not fixed and varies among subjects, whether CWGEE is still valid needs further investigation. In addition, the test for goodness of fit of CWGEE model with different correlation structures may deserve further investigation. In this thesis, we only consider the correlation structure within subject and ignore that within cluster. This may lose some information. How to choose the most informative correlation structure in CWGEE model is still an open question. These questions will be investigated elsewhere. Also, we are applying these methods to the dental study on periodontal disease, and then learn how to handle these in real dataset.

REFERENCES

- [1] D. Hedeker, *An Introduction To Growth Modeling*, Thousand Oaks CA: Sage Publications, 2004.
- [2] K. Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear model," *Biometrika*, vol. 73, pp. 13–22, 1986.
- [3] G. A. Ballinger, "Using generalized estimating equations for longitudinal data analysis," *Organizational Research Methods*, vol. 24, pp. 127–150, 2004.
- [4] J. S. Preisser and B. F. Qaqish, "Robust regression for clustered data with application to binary responses," *Biometrics*, vol. 55, pp. 574–579, 1999.
- [5] M. Crowder, "On the use of a working correlation matrix in using generalized linear models for repeated measures," *Biometrika*, vol. 4, pp. 407–410, 1995.
- [6] W. Mason, "Multilevel methods of statistical analysis," *International Encyclopedia of the Social and Behavioral Sciences*, , no. 34, 2001.
- [7] D. Hedeker and R. D. Gibbons, *Longitudinal data analysis*, John Wiley & Sons, Inc. Hoboken, New Jersey, 2006.
- [8] I. G. G. Kreft and J. De Leeuw, *Introducing Multilevel Modeling*, London: Sage, 1998.
- [9] H. Goldstein, W. Browne, and J. Rasbash, "Multilevel modelling of medical data," *Statistics in Medicine*, vol. 25, pp. 3291–3315, 2002.
- [10] S. Datta and G. A. Satten, "Rank-sum tests for clustered data," *Journal of the American Statistical Association*, vol. 100, pp. 908–915, 2005.
- [11] R. P. Tett, D. N. Jackson, and M. Rothstein, "Personality measures as predictors of job performance: A meta-analytic review," *Personnel Psychology*, vol. 44, pp. 703–741, 1991.
- [12] S. R. Searle, G. Casella, and C. E. McCulloch, *Variance Components*, New Jersey, Wiley-Interscience, 1992.
- [13] A. S. Bryk and S. W. Raudenbush, *Hierarchical linear models*, CA: Sage, 1992.
- [14] W. R. Gilks, C. C. Wang, B. Yvonnet, and P. Coursaget, "Random effect models for longitudinal data using gibbs sampling," *Statistical Science*, vol. 7, pp. 457–472, 1993.
- [15] M. Daniels and C. Gatsonis, "Hierarchical generalized linear models in the analysis of variations in health care utilization," *Journal of the American Statistical Association*, vol. 94, no. 445, 1999.
- [16] M. R. Steenbergen and B. S. Jones, "Modeling multilevel data structures," *American Journal of Political Science*, vol. 59, pp. 218–237, 2002.

- [17] H. Goldstein, *Multilevel statistical models (2nd ed.)*, New York, NY: Halsted Press, 1995.
- [18] X. J. Cong, G. Yin, and Y. Shen, "Marginal analysis of correlated failure time data with informative cluster sizes," *Biometrics*, vol. 63, pp. 663–672, 2007.
- [19] Higgins M., Province M., Heiss G., Eckfekt J., Ellison R. C., Folsom A. R., Rao D. C., Sprafka J. M., and Williams R., "Nhlbi family heart study: objectives and design," *Am J Epidemiol*, vol. 143, pp. 1219–1228, 1996.
- [20] J. Cai and B. F. Qaqish, "Marginal analysis for cluster-based case-control studies," *The Indian Journal Of Statistics*, vol. 63, pp. 326–337, 2001.
- [21] E. B. Hoffman, P. K. Sen, and C. R. Weinberg, "Within-cluster resampling," *Biometrika*, vol. 88, pp. 1121–1134, 2001.
- [22] J. M. Williamson, S. Datta, and G. A. Satten, "Marginal analyses of clustered data when cluster size is informative," *Biometrics*, vol. 59, pp. 36–42, 2003.
- [23] D. Follmann, M. Proschan, and E. Leifer, "Multiple outputation: Inference for complex clustered data by averaging analyses from independent data," *Biometrics*, vol. 59, pp. 420–429, 2003.
- [24] E. Benhin, J. N. K. Rao, and A. J. Scott, "Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes," *Biometrika*, vol. 92, pp. 435–450, 2005.
- [25] C. J. W. Zorn, "Generalized estimating equation models for correlated data: A review with applications," *American Journal of Political Science*, vol. 45, pp. 470–490, 2001.
- [26] R. W. M. Wedderburn, "Quasi-likelihood functions, generalized linear models, and the gauss-newton method," *Biometrika*, vol. 61, pp. 439–447, 1974.
- [27] P. McCullagh and J. A. Nelder, "Quasi-likelihood functions," *Annals of Statistics*, vol. 11, pp. 59–67, 1983.
- [28] G. Y. Yi and R. J. Cook, "Marginal methods for incomplete longitudinal data arising in clusters," *Journal of the American Statistical Association*, vol. 97, no. 460, 2002.
- [29] K. Y. Liang and S. L. Zeger, "On the use of concordant pairs in matched case-control studies," *Biometrics*, vol. 44, pp. 1145–1156, 1998.
- [30] N. R. Chaganty, "An alternative approach to the analysis of longitudinal data via generalized estimating equations," *Journal of Statistical Planning and Inference*, vol. 63, pp. 39–54, 1997.
- [31] J. Shults and N. R. Chaganty, "Analysis of serially correlated data using quasi-least squares," *Biometrics*, vol. 54, pp. 1622–1630, 1998.
- [32] N. R. Chaganty and J. Shults, "On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter," *Journal of Statistical Planning and Inference*, vol. 76, pp. 145–161, 1999.

- [33] J. N. K. Rao, A. J. Scott, and E. Benhin, "Analysing cluster-correlated data with nonignorable cluster sizes," *Biometrika*, vol. 92, pp. 435–450, 2005.
- [34] O. Gaudoin, B. Yang, and M. Xie, "A simple goodness-of-fit test for the power-law process, based on the duane plot," *IEEE Transactions On Relability*, vol. 52, no. 1, 2003.

Appendix: R code for the thesis

```
#####
#####      Clustered longitudinal data      #####
#####

install.packages("gee")
library("gee")
library(MASS)

loop<-1000

original<-matrix(0,nrow=loop,ncol=4)

beta.gee1<-beta.gee2<-beta.gee3<-original
se.gee11<-se.gee12<-se.gee21<-se.gee22<-se.gee31<-se.gee32<-original
initial<-matrix(0,nrow=loop,ncol=1)
msey.gee1<-msey.gee2<-msey.gee3<-correlation.gee1<-correlation.gee3<-initial
beta.wcr<-se.wcr<-original msey.wcr<-correlation.wcr<-initial
data1<-list(0)
mse<-rep(0,4)

#####
#####      Noninformative clustered longitudinal data      #####
#####      Generate sample datasets with 50 clusters      #####
#####

for(num in 1:loop){

  clustereffect<-rnorm(50,0,0.25)

  exposegroup<-c(rep(0,25),rep(1,25))

  efunction<-1.5

  cluster<-0

  for (i in 1:50){

    cluster[i]<-rpois(1,exp(efunction))+1}

  cluster<-as.integer(cluster)

  unexpose<-cluster[1:25]
  expose<-cluster[26:50]

  unexpose.effect<-clustereffect[1:25]
  expose.effect<-clustereffect[26:50]

  b <- c(0.5,1.5,0.75,0.8)

  N<-sum(unexpose,expose)

  size1<-size2<-pre1<-pre2<-0

  for(i in 1:length(unexpose)){

    size1<-c(size1,rep(i,3*unexpose[i]))

    clustersize1<-size1[-1]

    pre1<-c(pre1,rep(unexpose.effect[i],each=3*unexpose[i]))
```

```

        cluster.unexpose<-pre1[-1]}
for(i in 1:length(expose)){
    size2<-c(size2,rep(i+length(unexpose),3*expose[i]))
    clustersize2<-size2[-1]
    pre2<-c(pre2,rep(expose.effect[i],each=3*expose[i]))
    cluster.expose<-pre2[-1]}
data<-data.frame(subject=rep(1:N,each=3),time=rep(1:3,N),cluster=c(clustersize1,
clustersize2),predict=c(rep(0,3*sum(unexpose)),rep(1,3*sum(expose))),clustereffect=
rep(1,N*3),subjecteffect=rep(0,N*3),response=rep(0,N*3))
data$clustereffect<-c(cluster.unexpose,cluster.expose)
data$subjecteffect<-rep(rnorm(N,0,0.15),each=3)
data$response<-b[1]+b[2]*data$time+b[3]*data$predict+b[4]*data$predict*data$time+
data$clustereffect+data$subjecteffect+rnorm(N*3,0,0.01)
data1[[num]]<-data
}

#####
##### Noninformative clustered longitudinal data #####
##### Generate sample datasets with 500 clusters #####
#####
for (num in 1:loop){
    clustereffect<-rnorm(500,0,0.25)
    exposegroup<-c(rep(0,250),rep(1,250))
    efunction<-1.5
    cluster<-0
    for (i in 1:500){
        cluster[i]<-rpois(1,exp(efunction))+1}
    cluster<-as.integer(cluster)
    unexpose<-cluster[1:250]
    expose<-cluster[251:500]
    unexpose.effect<-clustereffect[1:250]
    expose.effect<-clustereffect[251:500]
    b <- c(0.5,1.5,0.75,0.8)
    N<-sum(unexpose,expose)
    size1<-size2<-pre1<-pre2<-0
    for(i in 1:length(unexpose)){
        size1<-c(size1,rep(i,3*unexpose[i]))
        clustersize1<-size1[-1]
        pre1<-c(pre1,rep(unexpose.effect[i],each=3*unexpose[i]))
        cluster.unexpose<-pre1[-1]}

```

```

for(i in 1:length(expose)){
  size2<-c(size2,rep(i+length(unexpose),3*expose[i]))
  clustersize2<-size2[-1]
  pre2<-c(pre2,rep(expose.effect[i],each=3*expose[i]))
  cluster.expose<-pre2[-1]}
data<-data.frame(subject=rep(1:N,each=3),time=rep(1:3,N),cluster=c(clustersize1,
clustersize2),predict=c(rep(0,3*sum(unexpose)),rep(1,3*sum(expose))),clustereffect=
rep(1,N*3),subjecteffect=rep(0,N*3),response=rep(0,N*3))
data$clustereffect<-c(cluster.unexpose,cluster.expose)
data$subjecteffect<-rep(rnorm(N,0,0.15),each=3)
data$response<-b[1]+b[2]*data$time+b[3]*data$predict+b[4]*data$predict*data$time+
data$clustereffect+data$subjecteffect+rnorm(N*3,0,0.01)
data1[[num]]<-data
}

```

```

#####
##### Generate sample datasets with 50 clusters #####
##### Cluster size related to clustereffect #####
#####

```

```

for (num in 1:loop){
  gama<-c(3,5)
  clustereffect<-rnorm(50,0,0.25)
  exposegroup<-c(rep(0,25),rep(1,25))
  efunction<-gama[1]+gama[2]*clustereffect
  cluster<-0
  for(i in 1:length(efunction)){
    cluster[i]<-rpois(1,exp(efunction[i]))+1}
  cluster<-as.integer(cluster)
  unexpose<-cluster[1:25]
  expose<-cluster[26:50]
  unexpose.effect<-clustereffect[1:25]
  expose.effect<-clustereffect[26:50]
  b <- c(0.5,1.5,0.75,0.8)
  N<-sum(unexpose,expose)
  size1<-size2<-pre1<-pre2<-0
  for(i in 1:length(unexpose)){
    size1<-c(size1,rep(i,3*unexpose[i]))
    clustersize1<-size1[-1]
    pre1<-c(pre1,rep(unexpose.effect[i],each=3*unexpose[i]))
    cluster.unexpose<-pre1[-1]}

```

```

for(i in 1:length(expose)){
  size2<-c(size2,rep(i+length(unexpose),3*expose[i]))
  clustersize2<-size2[-1]
  pre2<-c(pre2,rep(expose.effect[i],each=3*expose[i]))
  cluster.expose<-pre2[-1]}
data<-data.frame(subject=rep(1:N,each=3),time=rep(1:3,N),cluster=c(clustersize1,
clustersize2),predict=c(rep(0,3*sum(unexpose)),rep(1,3*sum(expose))),clustereffect=
rep(1,N*3),subjecteffect=rep(0,N*3),response=rep(0,N*3))
data$clustereffect<-c(cluster.unexpose,cluster.expose)
data$subjecteffect<-rep(rnorm(N,0,0.15),each=3)
data$response<-b[1]+b[2]*data$time+b[3]*data$predict+b[4]*data$predict*data$time+
data$clustereffect+data$subjecteffect+rnorm(N*3,0,0.01)
data1[[num]]<-data
}

```

```

#####
##### Generate sample datasets with 50 clusters #####
####Cluster sized related to clustereffect and exposed factor####
#####
for (num in 1:loop){
  gama<-c(3,5)
  clustereffect<-rnorm(50,0,0.25)
  exposegroup<-c(rep(0,25),rep(1,25))
  efunction<-gama[1]+gama[2]*clustereffect
  cluster<-0
  for (i in 1:length(efunction)){
    cluster[i]<-rpois(1,exp(efunction[i]-5*exposegroup[i]))+1}
  cluster<-as.integer(cluster)
  unexpose<-cluster[1:25]
  expose<-cluster[26:50]
  unexpose.effect<-clustereffect[1:25]
  expose.effect<-clustereffect[26:50]
  b <- c(0.5,1.5,0.75,0.8)
  N<-sum(unexpose,expose)
  size1<-size2<-pre1<-pre2<-0
  for(i in 1:length(unexpose)){
    size1<-c(size1,rep(i,3*unexpose[i]))
    clustersize1<-size1[-1]
    pre1<-c(pre1,rep(unexpose.effect[i],each=3*unexpose[i]))
    cluster.unexpose<-pre1[-1]}

```

```

    for(i in 1:length(expose)){
      size2<-c(size2,rep(i+length(unexpose),3*expose[i]))
      clustersize2<-size2[-1]
      pre2<-c(pre2,rep(expose.effect[i],each=3*expose[i]))
      cluster.expose<-pre2[-1]}
    data<-data.frame(subject=rep(1:N,each=3),time=rep(1:3,N),cluster=c(clustersize1,
    clustersize2),predict=c(rep(0,3*sum(unexpose)),rep(1,3*sum(expose))),clustereffect=
    rep(1,N*3),subjecteffect=rep(0,N*3),response=rep(0,N*3))
    data$clustereffect<-c(cluster.unexpose,cluster.expose)
    data$subjecteffect<-rep(rnorm(N,0,0.15),each=3)
    data$response<-b[1]+b[2]*data$time+b[3]*data$predict+b[4]*data$predict*data$time+
    data$clustereffect+data$subjecteffect+rnorm(N*3,0,0.01)
    data1[[num]]<-data
  }

#####
#####  Generate sample datasets with 50 clusters #####
#####  Cluster size related to exposed factor #####
#####  Drop clustereffect and subjecteffect #####
#####
#AR1 Structure

cormax.ar1<- function(alpha){
  n.max<-3
  cor.max<- diag(1,n.max)
  lowertri<- rep(0,0)
  for(j in (n.max-1):1){
    lowertri<- c(lowertri,1:j)
  }
  cor.max[lower.tri(cor.max)]<- alpha^lowertri
  cor.max[upper.tri(cor.max)]<- alpha^lowertri[length(lowertri):1]
  return(cor.max)
}

for (num in 1:loop){
  gama<--5
  exposegroup<-c(rep(0,25),rep(1,25))
  efunction<-gama*exposegroup
  cluster<-0
  for(i in 1:length(efunction)){

```



```

        cluster[i]<-rpois(1,exp(efunction[i]))+1}
cluster<-as.integer(cluster)
unexpose<-cluster[1:25]
expose<-cluster[26:50]
b <- c(0.5,1.5,0.75,0.8)
N<-sum(unexpose,expose)
size1<-size2<-0
for(i in 1:length(unexpose)){
    size1<-c(size1,rep(i,3*unexpose[i]))
    clustersize1<-size1[-1]}
for(i in 1:length(expose)){
    size2<-c(size2,rep(i+length(unexpose),3*expose[i]))
    clustersize2<-size2[-1]}
data<-data.frame(subject=rep(1:N,each=3),time=rep(1:3,N),cluster=c(clustersize1,
clustersize2),predict=c(rep(0,3*sum(unexpose)),rep(1,3*sum(expose))),response=rep(0,N*3))
residual<-mvrnorm(N,rep(0,3),0.5*cormax.ar1(0.9))
data$response<-b[1]+b[2]*data$time+b[3]*data$predict+b[4]*data$predict*data$time+
as.vector(t(residual))
data1[[num]]<-data
}

#####
##### Generate sample datasets with 50 clusters #####
##### Cluster size related to clustereffect #####
##### Drop subject effect #####
#####
#AR1 Structure

cormax.ar1<- function(alpha){
    n.max<-3
    cor.max<- diag(1,n.max)
    lowertri<- rep(0,0)
    for(j in (n.max-1):1){
        lowertri<- c(lowertri,1:j)
    }
    cor.max[lower.tri(cor.max)]<- alpha^lowertri
    cor.max[upper.tri(cor.max)]<- alpha^lowertri[length(lowertri):1]
    return(cor.max)
}

```

```

for (num in 1:loop){
  gama<-c(3,5)
  clustereffect<-rnorm(50,0,0.25)
  exposegroup<-c(rep(0,25),rep(1,25))
  efunction<-gama[1]+gama[2]*clustereffect
  cluster<-0
  for (i in 1:length(efunction)){
    cluster[i]<-rpois(1,exp(efunction[i]))+1}
  cluster<-as.integer(cluster)
  unexpose<-cluster[1:25]
  expose<-cluster[26:50]
  unexpose.effect<-clustereffect[1:25]
  expose.effect<-clustereffect[26:50]
  b <- c(0.5,1.5,0.75,0.8)
  N<-sum(unexpose,expose)
  size1<-size2<-pre1<-pre2<-0
  for(i in 1:length(unexpose)){
    size1<-c(size1,rep(i,3*unexpose[i]))
    clustersize1<-size1[-1]
    pre1<-c(pre1,rep(unexpose.effect[i],each=3*unexpose[i]))
    cluster.unexpose<-pre1[-1]}
  for(i in 1:length(expose)){
    size2<-c(size2,rep(i+length(unexpose),3*expose[i]))
    clustersize2<-size2[-1]
    pre2<-c(pre2,rep(expose.effect[i],each=3*expose[i]))
    cluster.expose<-pre2[-1]}
  data<-data.frame(subject=rep(1:N,each=3),time=rep(1:3,N),cluster=c(clustersize1,
clustersize2),predict=c(rep(0,3*sum(unexpose)),rep(1,3*sum(expose))),clustereffect=
rep(1,N*3),response=rep(0,N*3))
  data$clustereffect<-c(cluster.unexpose,cluster.expose)
  residual<-mvrnorm(N,rep(0,3),0.5*cormax.ar1(0.9))
  data$response<-b[1]+b[2]*data$time+b[3]*data$predict+b[4]*data$predict*data$time+
  data$clustereffect+as.vector(t(residual))
  data1[[num]]<-data
}

```

```

#####
##### Generate sample datasets with 50 clusters #####
#### Cluster size related to clustereffect and exposed factor ####
##### Drop subject effect #####

```

```
#####
#AR1 Structure

cormax.ar1<- function(alpha){
  n.max<-3
  cor.max<- diag(1,n.max)
  lowertri<- rep(0,0)
  for(j in (n.max-1):1){
    lowertri<- c(lowertri,1:j)
  }
  cor.max[lower.tri(cor.max)]<- alpha^lowertri
  cor.max[upper.tri(cor.max)]<- alpha^lowertri[length(lowertri):1]
  return(cor.max)
}

for (num in 1:loop){
  gama<-c(3,5)
  clustereffect<-rnorm(50,0,0.25)
  exposegroup<-c(rep(0,25),rep(1,25))
  efunction<-gama[1]+gama[2]*clustereffect
  cluster<-0
  for (i in 1:length(efunction)){
    cluster[i]<-rpois(1,exp(efunction[i]-5*exposegroup[i]))+1}
  cluster<-as.integer(cluster)
  unexpose<-cluster[1:25]
  expose<-cluster[26:50]
  unexpose.effect<-clustereffect[1:25]
  expose.effect<-clustereffect[26:50]
  b <- c(0.5,1.5,0.75,0.8)
  N<-sum(unexpose,expose)
  size1<-size2<-pre1<-pre2<-0
  for(i in 1:length(unexpose)){
    size1<-c(size1,rep(i,3*unexpose[i]))
    clustersize1<-size1[-1]
    pre1<-c(pre1,rep(unexpose.effect[i],each=3*unexpose[i]))
    cluster.unexpose<-pre1[-1]}
  for(i in 1:length(expose)){
    size2<-c(size2,rep(i+length(unexpose),3*expose[i]))
    clustersize2<-size2[-1]
    pre2<-c(pre2,rep(expose.effect[i],each=3*expose[i]))
  }
}
```

```

cluster.expose<-pre2[-1]}

data<-data.frame(subject=rep(1:N,each=3),time=rep(1:3,N),cluster=c(cluster.size1,
cluster.size2),predict=c(rep(0,3*sum(unexpose)),rep(1,3*sum(expose))),clustereffect=
rep(1,N*3),response=rep(0,N*3))

data$clustereffect<-c(cluster.unexpose,cluster.expose)

residual<-mvrnorm(N,rep(0,3),0.5*cormax.ar1(0.9))

data$response<-b[1]+b[2]*data$time+b[3]*data$predict+b[4]*data$predict*data$time+
data$clustereffect+as.vector(t(residual))

data1[[num]]<-data
}

#####
#####      GEE model with different structures      #####
#####
## GEE model: Exchangeable correlation structure
se.mean.gee1<-se.mean.gee12<-rep(0,4)
n.gee1<-rep(0,4)

for (num in 1:loop){
  CI<-matrix(0,nrow=4,ncol=2)

  fit.gee1<-gee(response~time+predict+time*predict,data=data1[[num]],id=subject,family=
gaussian,corstr="exchangeable")

  mse.gee1[num,]<-sum((fit.gee1$fitted.values-data1[[num]]$response)^2)/(length(data1[[num]]$
response)-4)

  beta.gee1[num,]<-fit.gee1$coefficient

  correlation.gee1[num,]<-fit.gee1$working.correlation[1,2]

  for(m in 1:4){se.gee11[num,][m]<-fit.gee1$naive.variance[m,m]}

  for(m in 1:4){se.gee12[num,][m]<-fit.gee1$robust.variance[m,m]}

  for(m in 1:4){
    CI[m,]<-beta.gee1[num,][m]+c(-1,1)*1.96*se.gee12[num,][m]^0.5
    ifelse(CI[m,][1]<b[m]& b[m]<CI[m,][2],n.gee1[m]<-n.gee1[m]+1,n.gee1[m])}
  }

beta.mean.gee1<-apply(beta.gee1,2,mean)

for(i in 1:4){
  se.mean.gee11[i]<-sqrt(sum(se.gee11[,i]/(loop^2)))
  se.mean.gee12[i]<-sqrt(sum(se.gee12[,i]/(loop^2)))}

for(i in 1:4){
  mse[i]<-sum((beta.gee1[,i]-b[i])^2)/loop}

coverage<-n.gee1/loop

correlation.mean.gee1<-apply(correlation.gee1,2,mean)

```

```

msey.mean.gee1<-apply(msey.gee1,2,mean)
msey.sd.gee1<-apply(msey.gee1,2,sd)
beta.sd.gee1<-apply(beta.gee1,2,sd)
correlation.sd.gee1<-apply(correlation.gee1,2,sd)

## GEE model: Independent correlation structure
se.mean.gee21<-se.mean.gee22<-rep(0,4)
n.gee2<-rep(0,4)

for (num in 1:loop){
  CI<-matrix(0,nrow=4,ncol=2)
  fit.gee2<-gee(response~time+predict+time*predict,data=data1[[num]],id=subject,family=
    gaussian,corstr="independence")
  msey.gee2[num,<-sum((fit.gee2$fitted.values-data1[[num]]$response)^2)/(length(data1[[num]]$
    response)-4)
  beta.gee2[num,<-fit.gee2$coefficient
  for(m in 1:4){se.gee21[num,][m]<-fit.gee2$naive.variance[m,m]}
  for(m in 1:4){se.gee22[num,][m]<-fit.gee2$robust.variance[m,m]}
  for(m in 1:4){
    CI[m,<-beta.gee2[num,][m]+c(-1,1)*1.96*se.gee22[num,][m]^0.5
    ifelse(CI[m,][1]<b[m]& b[m]<CI[m,][2],n.gee2[m]<-n.gee2[m]+1,n.gee2[m])}
  }
beta.mean.gee2<-apply(beta.gee2,2,mean)
for(i in 1:4){
  se.mean.gee21[i]<-sqrt(sum(se.gee21[,i]/(loop^2)))
  se.mean.gee22[i]<-sqrt(sum(se.gee22[,i]/(loop^2)))}
msey.mean.gee2<-apply(msey.gee2,2,mean)
msey.sd.gee2<-apply(msey.gee2,2,sd)
for(i in 1:4){
  mse[i]<-sum((beta.gee2[,i]-b[i])^2)/loop}
coverage<-n.gee2/loop
beta.sd.gee2<-apply(beta.gee2,2,sd)
correlation.mean.gee1<-apply(correlation.gee1,2,mean)

## GEE model: AR1 correlation structure
se.mean.gee31<-se.mean.gee32<-rep(0,4)
n.gee3<-rep(0,4)

for (num in 1:loop){
  CI<-matrix(0,nrow=4,ncol=2)

```

```

fit.gee3<-gee(response~time+predict+time*predict,data=data1[[num]],id=subject,family=
gaussian,corstr="AR-M")
msey.gee3[num,]<-sum((fit.gee3$fitted.values-data1[[num]]$response)^2)/(length(data1[[num]]$
response)-4)
beta.gee3[num,]<-fit.gee3$coefficient
correlation.gee3[num,]<-fit.gee3$working.correlation[1,2]
for(m in 1:4){se.gee31[num,][m]<-fit.gee3$naive.variance[m,m]}
for(m in 1:4){se.gee32[num,][m]<-fit.gee2$robust.variance[m,m]}
for(m in 1:4){
  CI[m,]<-beta.gee3[num,][m]+c(-1,1)*1.96*se.gee32[num,][m]^0.5
  ifelse(CI[m,][1]<b[m]& b[m]<CI[m,][2],n.gee3[m]<-n.gee3[m]+1,n.gee3[m])}
}
beta.mean.gee3<-apply(beta.gee3,2,mean)
for(i in 1:4){
  se.mean.gee31[i]<-sqrt(sum(se.gee31[,i]/(loop^2)))
  se.mean.gee32[i]<-sqrt(sum(se.gee32[,i]/(loop^2)))}
msey.mean.gee3<-apply(msey.gee3,2,mean)
msey.sd.gee3<-apply(msey.gee3,2,sd)
for(i in 1:4){
  mse[i]<-sum((beta.gee3[,i]-b[i])^2)/loop}
coverage<-n.gee3/loop
beta.sd.gee3<-apply(beta.gee3,2,sd)
correlation.mean.gee3<-apply(correlation.gee3,2,mean)
correlation.sd.gee3<-apply(correlation.gee3,2,sd)

#####
#####      WCR model with different structures      #####
#####

n.wcr<-rep(0,4)
for (num in 1:loop){
  beta<-matrix(0,nrow=5000,ncol=4)
  var<-matrix(0,nrow=5000,ncol=4)
  corr<-msey<-matrix(0,nrow=5000)
  CI<-matrix(0,nrow=4,ncol=2)
  t<-length(c(unexpose,expose))
  for(lloop in 1:5000){
    dataloop<-data.frame(subj=rep(0,3*t),predict=rep(0,3*t),time=rep(0,3*t),cluster=rep
      (0,3*t),resp=rep(0,3*t))
    for (i in 1:t){

```

```

response<-data1[[num]]$response[data1[[num]]$cluster==i]
sample<-sample(response,1,replace=TRUE)
subj<-data1[[num]]$subject[data1[[num]]$cluster==i&data1[[num]]$response[data1[[num]]$
cluster==i]==sample]
for(j in 1:3){
  observ<-0
  observ<-data1[[num]]$response[data1[[num]]$subject==subj][j]
  dataloop$resp[j+3*(i-1)]<-observ
  dataloop$subj[j+3*(i-1)]<-subj
  dataloop$time[j+3*(i-1)]<-data1[[num]]$time[data1[[num]]$subject==subj][j]
  dataloop$predict[j+3*(i-1)]<-data1[[num]]$predict[data1[[num]]$subject==subj][j]
  dataloop$cluster[j+3*(i-1)]<-i
}
}
attach(dataloop)
###fit.wcr<-gee(resp~time+predict+time*predict,data=dataloop,id=subj,family=gaussian,
corstr="exchangeable")###
###fit.wcr<-gee(resp~time+predict+time*predict,data=dataloop,id=subj,family=gaussian,
corstr="independence")###
fit.wcr<-gee(resp~time+predict+time*predict,data=dataloop,id=subj,family=gaussian,
corstr="AR-M")
mseyl[loop,]<-sum((fit.wcr$fitted.values-dataloop$resp)^2)/(length(dataloop$resp)-4)
for(m in 1:4){var[lloop,][m]<-fit.wcr$robust.variance[m,m]}
beta[lloop,]<-fit.wcr$coefficients
corr[lloop,]<-fit.wcr$working.correlation[1,2]
detach(dataloop)
}
beta.wcr[num,]<-apply(beta,2,mean)
var1<-matrix(0,nrow=4,ncol=4)
for (q in 1:5000){
  part<-(as.matrix(beta[q,]-beta.wcr[num,]))%*%t(as.matrix(beta[q,]-beta.wcr[num,]))
  var1<-var1+part}
se.wcr[num,]<-(apply(var,2,mean)-diag(var1/5000)^0.5
mseyl.wcr[num,]<-apply(mseyl,2,mean)
correlation.wcr[num,]<-apply(corr,2,mean) for(m in 1:4){
  CI[m,]<-beta.wcr[num,][m]+c(-1,1)*1.96*se.wcr[num,][m]
  ifelse(CI[m,][1]<b[m]& b[m]<CI[m,][2],n.wcr[m]<-n.wcr[m]+1,n.wcr[m])}
}
beta.mean.wcr<-apply(beta.wcr,2,mean)
se.mean.wcr<-apply(na.omit(se.wcr),2,mean)

```



```

msey.mean.wcr<-apply(msey.wcr,2,mean)
msey.sd.wcr<-apply(msey.wcr,2,sd) for(i in 1:4){
  mse[i]<-sum((beta.wcr[,i]-b[i])^2)/loop}
coverage<-n.wcr/loop
coverage[2]<-n.wcr[2]/length(na.omit(se.wcr[,2]))
coverage[4]<-n.wcr[4]/length(na.omit(se.wcr[,4]))
beta.sd.wcr<-apply(beta.wcr,2,sd)
correlation.mean.wcr<-apply(correlation.wcr,2,mean)
correlation.sd.wcr<-apply(correlation.wcr,2,sd)

#####
##### CWGEE model with different structures #####
#####

##### Getting Info from the Data #####
cluster.size<- function(id){
  clid<- unique(id)
  m<- length(unique(id))
  n<- rep(0,m)
  autotime<- rep(0,0)
  for(i in 1:m){
    n[i]<- length(which(id==clid[i]))/3
    autotime<- c(autotime,rep(1:n[i],each=3))
  }
  id<- rep(1:m,n*3)
  return(list(m=m,n=n,id=id,autotime=autotime))
}

##### Calculating Residuals #####
residual<- function(x,y,beta,family){
  x<- as.matrix(x)
  y<- as.matrix(y)
  beta<- as.matrix(beta)
  u<- switch(family,
    gaussian=x%*%beta,
    binomial=exp(x%*%beta)/(1+exp(x%*%beta)),
    poisson=exp(x%*%beta)
  )
  h<- switch(family,

```

```

        gaussian=1,
        binomial=u*(1-u),
        poisson=u
    )

    return((y-u)/sqrt(h))
}

##### Correlation Matrix #####
#AR1 Structure
cormax.ar1<- function(alpha){
    n.max<-3
    cor.max<- diag(1,n.max)
    lowertri<- rep(0,0)
    for(j in (n.max-1):1){
        lowertri<- c(lowertri,1:j)
    }
    cor.max[lower.tri(cor.max)]<- alpha^lowertri
    cor.max[upper.tri(cor.max)]<- alpha^lowertri[length(lowertri):1]
    return(cor.max)
}

#Exchangeable Structure
cormax.exch<- function(alpha){
    n.max<- 3
    cor.max<- diag(1,n.max)
    cor.max[lower.tri(cor.max)]<- rep(alpha,n.max*(n.max-1)/2)
    cor.max[upper.tri(cor.max)]<- rep(alpha,n.max*(n.max-1)/2)
    return(cor.max)
}

#Independent Structure
cormax.ind<-function(n){
    matrix<-matrix(0,nrow=n,ncol=n)
    diag(matrix)<-rep(1,n)
    return(matrix)
}

##### estimate beta #####
gee.fixed<-function(data,alpha,correlation){
    var<- switch(correlation,

```

```

        ari=cormax.ari(alpha),
        exchangeable=cormax.exch(alpha),
        independence=cormax.ind(3),
    )

    beta<-rep(0,4)
    cluster<-cluster.size(data$cluster)
    group<-cluster$n
    size<-length(group)
    step21<-matrix(0,nrow=4,ncol=1)
    step22<-matrix(0,nrow=4,ncol=4)
    for (i in 1:size){
        step11<-matrix(0,nrow=4,ncol=1)
        step12<-matrix(0,nrow=4,ncol=4)
        for( j in unique(data$subject[data$cluster==i])){
            p<-as.matrix(data$predict[data$subject==j])
            t<-as.matrix(data$time[data$subject==j])
            pt<-p*t
            cons<-as.matrix(c(1,1,1))
            y<-as.matrix(data$response[data$subject==j])
            covariate<-cbind(cons,t,p,pt)
            xy<-t(covariate)%*%solve(var)%*%y
            xx<-t(covariate)%*%solve(var)%*%covariate
            step11<-step11+xy
            step12<-step12+xx
        }
        step11<-step11/group[i]
        step12<-step12/group[i]
        step21<-step21+step11
        step22<-step22+step12
    }

    beta<-as.vector(solve(step22,step21))
    return(beta)
}

##### estimate the robust variance of beta #####
beta.robust.sd<-function(data,alpha,beta,correlation){
    var<- switch(correlation,
        ari=cormax.ari(alpha),
        exchangeable=cormax.exch(alpha),
        independence=cormax.ind(3),

```

```

    )

    cov.beta<-matrix(0,nrow=4,ncol=4)
    cluster<-cluster.size(data$cluster)
    group<-cluster$n
    size<-length(group)
    step21<-matrix(0,nrow=4,ncol=4)
    step31<-matrix(0,nrow=4,ncol=4)
    step22<-matrix(0,nrow=4,ncol=4)
    for (i in 1:size){
        step11<-matrix(0,nrow=4,ncol=1)
        step12<-matrix(0,nrow=4,ncol=4)
        for( j in unique(data$subject[data$cluster==i])){
            p<-as.matrix(data$predict[data$subject==j])
            t<-as.matrix(data$time[data$subject==j])
            pt<-p*t
            cons<-as.matrix(c(1,1,1))
            y<-as.matrix(data$response[data$subject==j])
            covariate<-cbind(cons,t,p,pt)
            xy<-t(covariate)%*%solve(var)%*%(y-covariate)%*%beta)
            xx<-t(covariate)%*%solve(var)%*%covariate
            step11<-step11+xy
            step12<-step12+xx
        }
        step21<-(step11/group[i])%*%t(step11/group[i])
        step12<-step12/group[i]
        step31<-step31+step21
        step22<-step22+step12
    }

    cov.beta<-solve(step22)%*%(step31)%*%solve(step22)
    return(diag(cov.beta))
}

##### estimate the naive variance of beta #####
beta.naive.sd<-function(data,alpha,beta,correlation){
    var<- switch(correlation,
        ari=cormax.ari(alpha),
        exchangeable=cormax.exch(alpha),
        independence=cormax.ind(3),
    )
    cov.beta<-matrix(0,nrow=4,ncol=4)

```

```

cluster<-cluster.size(data$cluster)
group<-cluster$n
size<-length(group)
step22<-matrix(0,nrow=4,ncol=4)
for (i in 1:size){
  step12<-matrix(0,nrow=4,ncol=4)
  for( j in unique(data$subject[data$cluster==i])){
    p<-as.matrix(data$predict[data$subject==j])
    t<-as.matrix(data$time[data$subject==j])
    pt<-p*t
    cons<-as.matrix(c(1,1,1))
    covariate<-cbind(cons,t,p,pt)
    xx<-t(covariate)%*%solve(var)%*%covariate
    step12<-step12+xx
  }
  step12<-step12/group[i]
  step22<-step22+step12
}
cov.beta<- solve(step22)
return(diag(cov.beta))
}

beta.ar<-beta.exch<-original
beta.robust.var.ar<-beta.robust.var.exch<-original
beta.naive.var.ar<-beta.naive.var.exch<-original
beta.robust.sd.mean.ar<-beta.robust.sd.mean.exch<-rep(0,4)
beta.naive.sd.mean.ar<-beta.naive.sd.mean.exch<-rep(0,4)
mse.cwgee.ar<-mse.cwgee.exch<-initial
alpha.one.ar<-alpha.one.exch<-initial
alpha.two.ar<-alpha.two.exch<-initial mse.ar<-mse.exch<-rep(0,4)
n.ar<-n.exch<-rep(0,4)
for (t in 1:loop){
  CI.ar<-CI.exch<-matrix(0,nrow=4,ncol=2)
  ar<-qls(data1[[t]],family="gaussian",correlation="ar1")
  exch<-qls(data1[[t]],family="gaussian",correlation="exchangeable")
  beta.robust.var.ar[t,<-beta.robust.sd(data1[[t]],alpha=ar$alpha,beta=ar$beta,
  correlation="ar1")
  beta.robust.var.exch[t,<-beta.robust.sd(data1[[t]],alpha=exch$alpha,beta=exch$beta,
  correlation="exchangeable")
  beta.naive.var.ar[t,<-beta.naive.sd(data1[[t]],alpha=ar$alpha,beta=ar$beta,

```

```

correlation="ar1")
beta.naive.var.exch[t,]<-beta.naive.sd(data1[[t]],alpha=exch$alpha,beta=exch$beta,
correlation="exchangeable")
beta.ar[t,]<-ar$beta
beta.exch[t,]<-exch$beta
mse.cwgee.ar[t,]<-sum(ar$residual^2)/(length(data1[[t]]$response)-4)
mse.cwgee.exch[t,]<-sum(exch$residual^2)/(length(data1[[t]]$response)-4)
alpha.one.ar[t,]<-ar$alpha
alpha.two.ar[t,]<-ar1.two(alpha.one.ar[t,])
alpha.one.exch[t,]<-exch$alpha
alpha.two.exch[t,]<-exch.two(alpha.one.exch[t,])
for(m in 1:4){
  CI.ar[m,]<-beta.ar[t,][m]+c(-1,1)*1.96*beta.robust.var.ar[t,][m]^0.5
  ifelse(CI.ar[m,][1]<b[m]&b[m]<CI.ar[m,][2],n.ar[m]<-n.ar[m]+1,n.ar[m])
  CI.exch[m,]<-beta.exch[t,][m]+c(-1,1)*1.96*beta.robust.var.exch[t,][m]^0.5
  ifelse(CI.exch[m,][1]<b[m]&b[m]<CI.exch[m,][2],n.exch[m]<-n.exch[m]+1,n.exch[m])}
}
beta.mean.ar<-apply(beta.ar,2,mean)
beta.mean.exch<-apply(beta.exch,2,mean)
for (i in 1:4){
  beta.robust.sd.mean.ar[i]<-round(sqrt(sum(beta.robust.var.ar[,i]/(loop^2))),6)}
for (i in 1:4){
  beta.robust.sd.mean.exch[i]<-round(sqrt(sum(beta.robust.var.exch[,i]/(loop^2))),6)}
for (i in 1:4){
  beta.naive.sd.mean.ar[i]<-round(sqrt(sum(beta.naive.var.ar[,i]/(loop^2))),6)}
for (i in 1:4){
  beta.naive.sd.mean.exch[i]<-round(sqrt(sum(beta.naive.var.exch[,i]/(loop^2))),6)}
coverage.ar<-n.ar/loop
coverage.exch<-n.exch/loop
beta.sd.ar<-apply(beta.ar,2,sd)
beta.sd.exch<-apply(beta.exch,2,sd)
for(i in 1:4){
  mse.ar[i]<-(sum((beta.ar[,i]-b[i])^2)/loop)^0.5}
for(i in 1:4){
  mse.exch[i]<-(sum((beta.exch[,i]-b[i])^2)/loop)^0.5}
mse.mean.cwgee.ar<-apply(mse.cwgee.ar,2,mean)
mse.mean.cwgee.exch<-apply(mse.cwgee.exch,2,mean)
alpha.one.mean.ar<-apply(alpha.one.ar,2,mean)
alpha.one.mean.exch<-apply(alpha.one.exch,2,mean)
alpha.one.sd.ar<-apply(alpha.one.ar,2,sd)

```

```

alpha.one.sd.exch<-apply(alpha.one.exch,2,sd)
alpha.two.sd.ar<-apply(alpha.two.ar,2,sd)
alpha.two.sd.exch<-apply(alpha.two.exch,2,sd)
alpha.two.mean.ar<-apply(alpha.two.ar,2,mean)
alpha.two.mean.exch<-apply(alpha.two.exch,2,mean)

#####
#####          Power test for CWGEE model          #####
#####
b1<-seq(-1,1,by=0.04)
b3<-seq(-1,1,by=0.04)
power.ar.1<-power.ar.2<-power.exch.1<-power.exch.2<-rep(0,length(b3))
tstat.ar.1<-tstat.ar.2<-tstat.exch.1<-tstat.exch.2<- rep(0,loop)
se.ar.1<-se.ar.2<-se.exch.1<-se.exch.2<- rep(0,length(b3))
CI.ar.1<-CI.ar.2<-CI.exch.1<-CI.exch.2<-matrix(0,nrow=length(b3),ncol=2)
for(eta in 1:length(b3)){
  b <- c(b1[eta],1.5,b3[eta],0.8)
  data1<-list(0)
  original<-matrix(0,nrow=loop,ncol=4)
  beta.ar<-beta.exch<-original
  beta.robust.sd.ar<-beta.robust.sd.exch<-original
  unexpose<-expose<-matrix(0,nrow=loop,ncol=25)
  for (num in 1:loop){
    gama<-c(3,5)
    clustereffect<-rnorm(50,0,0.25)
    exposegroup<-c(rep(0,25),rep(1,25))
    efunction<-gama[1]+gama[2]*clustereffect
    cluster<-0
    for (i in 1:length(efunction)){
      cluster[i]<-rpois(1,exp(efunction))+1}
    cluster<-as.integer(cluster)
    unexpose[num,]<-cluster[1:25]
    expose[num,]<-cluster[26:50]
    unexpose.effect<-clustereffect[1:25]
    expose.effect<-clustereffect[26:50]
    N<-sum(unexpose[num,],expose[num,]) size1<-size2<-pre1<-pre2<-0
    for(i in 1:length(unexpose[num,])){
      size1<-c(size1,rep(i,3*unexpose[num,][i]))
      clustersize1<-size1[-1]
      pre1<-c(pre1,rep(unexpose.effect[i],each=3*unexpose[num,][i]))
    }
  }
}

```

```

        cluster.unexpose<-pre1[-1]}
    for(i in 1:length(expose[num,])){
        size2<-c(size2,rep(i+length(unexpose[num,]),3*expose[num,][i]))
        clustersize2<-size2[-1]
        pre2<-c(pre2,rep(expose.effect[i],each=3*expose[num,][i]))
        cluster.expose<-pre2[-1]}

    data<-data.frame(subject=rep(1:N,each=3),time=rep(1:3,N),cluster=c(clustersize1,
    clustersize2),predict=c(rep(0,3*sum(unexpose[num,])),rep(1,3*sum(expose[num,]))),
    clustereffect=rep(0,N*3),
    subjecteffect=rep(0,N*3),response=rep(0,N*3))
    data$clustereffect<-c(cluster.unexpose,cluster.expose)
    data$subjecteffect<-rep(rnorm(N,0,0.15),each=3)
    data$response<-b[1]+b[2]*data$time+b[3]*data$predict+b[4]*data$predict*data$time+
    data$clustereffect+data$subjecteffect+rnorm(N*3,0,0.01)
    data1[[num]]<-data
  }

#### For the AR1 correlation structure ####
n.ar.1<-n.ar.2<-0
for (t in 1:loop){
  ar<-qls(data1[[t]],family="gaussian",correlation="ar1")
  ifelse(sum(!is.na(ar$beta))==4, beta.ar[t,<-ar$beta, beta.ar[t,<-beta.ar[t-1,])
  beta.robust.sd.ar[t,<-beta.robust.sd(data1[[t]],alpha=ar$alpha,beta=beta.ar[t,],
  correlation="ar1")^0.5
  tstat.ar.1[t]<-beta.ar[t,][3]/beta.robust.sd.ar[t,][3]
  tstat.ar.2[t]<-beta.ar[t,][1]/beta.robust.sd.ar[t,][1]
  ifelse(abs(tstat.ar.1[t])>1.96,n.ar.1<-n.ar.1+1,n.ar.1)
  ifelse(abs(tstat.ar.2[t])>1.96,n.ar.2<-n.ar.2+1,n.ar.2)
}

#### For the exchangeable correlation structure ####
n.exch.1<-n.exch.2<-0
for (t in 1:loop){
  exch<-qls(data1[[t]],family="gaussian",correlation="exchangeable")
  ifelse(sum(!is.na(exch$beta))==4, beta.exch[t,<-exch$beta, beta.exch[t,<-beta.exch[t-1,])
  beta.robust.sd.exch[t,<-beta.robust.sd(data1[[t]],alpha=exch$alpha,beta=beta.exch[t,],
  correlation="exchangeable")^0.5
  tstat.exch.1[t]<-beta.exch[t,][3]/beta.robust.sd.exch[t,][3]
  tstat.exch.2[t]<-beta.exch[t,][1]/beta.robust.sd.exch[t,][1]
  ifelse(abs(tstat.exch.1[t])>12.706,n.exch.1<-n.exch.1+1,n.exch.1)
  ifelse(abs(tstat.exch.2[t])>12.706,n.exch.2<-n.exch.2+1,n.exch.2)
}

```



```

##### Get the power for AR1 correlation structure #####
power.ar.1[eta]<-n.ar.1/loop power.ar.2[eta]<-n.ar.2/loop
se.ar.1[eta]<-(power.ar.1[eta]*(1-power.ar.1[eta])/loop)^0.5
se.ar.2[eta]<-(power.ar.2[eta]*(1-power.ar.2[eta])/loop)^0.5
CI.ar.1[eta,]<-power.ar.1[eta]+c(-1,1)*1.96*se.ar.1[eta]
CI.ar.2[eta,]<-power.ar.2[eta]+c(-1,1)*1.96*se.ar.2[eta]

##### Get the power for Exchangeable correlation structure #####
power.exch.1[eta]<-n.exch.1/loop power.exch.2[eta]<-n.exch.2/loop
se.exch.1[eta]<-(power.exch.1[eta]*(1-power.exch.1[eta])/loop)^0.5
se.exch.2[eta]<-(power.exch.2[eta]*(1-power.exch.2[eta])/loop)^0.5
CI.exch.1[eta,]<-power.exch.1[eta]+c(-1,1)*1.96*se.exch.1[eta]
CI.exch.2[eta,]<-power.exch.2[eta]+c(-1,1)*1.96*se.exch.2[eta] }

##Plots of the power options(width=60)
plot(b1,power.ar.1,type="n",xlim=c(-1,1),ylim=c(0,1),col="blue",xlab="CWGEE:beta0",
ylab="power.cwgee",main="power plot for AR-M and Exchangeable
correlation structures") lines(b1,power.ar.1,type="l",col="black")
lines(b1,power.exch.1,type="l",col="green")
points(b1,power.ar.1,pch=20,cex=1.2)
points(b1,power.exch.1,pch=23,cex=1.2,col="red")
abline(h=0.05,col="gray60",lty=3)
legend(-1,0.4,legend=c("Exchangeable", "AR-M"),pch=c(20,23),col=c("black", "red"))
text(0.2,min(power.ar.2)-0.03,"0.059
CI:(0.044,0.074)",col="blue",cex=0.8)

plot(b3,power.ar.2,type="n",xlim=c(-1,1),ylim=c(0,1),col="blue",xlab="CWGEE:beta2",
ylab="power.cwgee",main="power plot for AR-M and Exchangeable
correlation structures")
lines(b3,power.ar.2,type="l",col="black")
lines(b3,power.exch.2,type="l",col="green")
points(b3,power.ar.2,pch=20,cex=1.2)
points(b3,power.exch.2,pch=23,cex=1.2,col="red")
abline(h=0.05,col="gray60",lty=3)
legend(-1,0.4,legend=c("Exchangeable", "AR-M"),pch=c(20,23),col=c("black", "red"))
text(0.2,min(power.ar.2)-0.03,"0.059 CI:(0.044,0.074)",col="blue",cex=0.8)

#####
##### Probability-Probability Plots #####
#####

```

```

### GEE model with exchangeable structure
alpha<-seq(0.01,0.85,by=0.05)
coverage.gee<-matrix(0,nrow=length(alpha),ncol=4)

for (t in 1:length(alpha)){
  n.gee1<-rep(0,4)
  for (num in 1:loop){
    CI<-matrix(0,nrow=4,ncol=2)
    fit.gee1<-gee(response~time+predict+time*predict,data=data1[[num]],id=subject,
    family=gaussian,corstr="exchangeable")
    beta.gee1[num,]<-fit.gee1$coefficient
    correlation.gee1[num,]<-fit.gee1$working.correlation[1,2]
    for(m in 1:4){
      se.gee12[num,][m]<-fit.gee1$robust.variance[m,m]}
    for(n in 1:4){
      CI[n,]<-beta.gee1[num,][n]+c(-1,1)*qnorm(1-alpha[t]/2)*se.gee12[num,][n]^0.5}
    for(p in 1:4){
      ifelse(CI[p,][1]<b[p]& b[p]<CI[p,][2],n.gee1[p]<-n.gee1[p]+1,n.gee1[p])}
    }
    coverage.gee[t,]<-n.gee1/loop
  }
}

### WCR model with Exchangeable structure
alpha<-seq(0.01,0.85,by=0.05)
coverage.wcr<-matrix(0,nrow=length(alpha),ncol=4)

for (k in 1:length(alpha)){
  n.wcr<-rep(0,4)
  for (num in 1:loop){
    beta<-matrix(0,nrow=5000,ncol=4)
    var<-matrix(0,nrow=5000,ncol=4)
    CI<-matrix(0,nrow=4,ncol=2)
    t<-length(c(unexpose,expose))
    for(lloop in 1:5000){
      dataloop <- data.frame(subj=rep(0,3*t),predict=rep(0,3*t),time=rep(0,3*t),
      cluster=rep(0,3*t),resp=rep(0,3*t))
      for (i in 1:t){
        response<-data1[[num]]$response[data1[[num]]$cluster==i]
        sample<-sample(response,1,replace=TRUE)
        subj<-data1[[num]]$subject[data1[[num]]$cluster==i&data1[[num]]$response[data1

```

```

[[num]]$cluster==i]==sample]
for(j in 1:3){
  observ<-0
  observ<-data1[[num]]$response[data1[[num]]$subject==subj][j]
  dataloop$resp[j+3*(i-1)]<-observ
  dataloop$subj[j+3*(i-1)]<-subj
  dataloop$time[j+3*(i-1)]<-data1[[num]]$time[data1[[num]]$subject==subj][j]
  dataloop$predict[j+3*(i-1)]<-data1[[num]]$predict[data1[[num]]$subject==subj][j]
  dataloop$cluster[j+3*(i-1)]<-i
}
}

attach(dataloop)
fit.wcr<-gee(resp~time+predict+time*predict,data=dataloop,id=subj,family=gaussian,
corstr="exchangeable")
for(m in 1:4){var[lloop,][m]<-fit.wcr$robust.variance[m,m]}
beta[lloop,]<-fit.wcr$coefficient
detach(dataloop)
}

beta.wcr[num,]<-apply(beta,2,mean)
var1<-matrix(0,ncol=4,nrow=4)
for (q in 1:5000){
  part<-(as.matrix(beta[q,]-beta.wcr[num,]))%*%t(as.matrix(beta[q,]-beta.wcr[num,]))
  var1<-var1+part}
se.wcr[num,]<-(apply(var,2,mean)-diag(var1/5000)^0.5
for(n in 1:4){
  CI[n,]<-beta.wcr[num,][n]+c(-1,1)*qnorm(1-alpha[k]/2)*se.wcr[num,][n]}
for(p in 1:4){
  ifelse(CI[p,][1]<b[p]& b[p]<CI[p,][2],n.wcr[p]<-n.wcr[p]+1,n.wcr[p])}
}

coverage.wcr[k,]<-n.wcr/loop
coverage.wcr[k,][2]<-n.wcr[2]/length(na.omit(se.wcr[,2]))
coverage.wcr[k,][4]<-n.wcr[4]/length(na.omit(se.wcr[,4]))
}

### CWGEE model with AR-M and Exchangeable structures
loop<-1000
original<-matrix(0,nrow=loop,ncol=4)
beta.ar<-beta.exch<-original
beta.robust.var.ar<-beta.robust.var.exch<-original
beta.robust.sd.mean.ar<-beta.robust.sd.mean.exch<-rep(0,4)

```

```

alpha<-seq(0.01,0.85,by=0.05)
coverage.ar<-coverage.exch<-matrix(0,nrow=length(alpha),ncol=4)
for (t in 1:length(alpha)){
  n.ar<-n.exch<-rep(0,4)
  for (num in 1:loop){
    CI.ar<-CI.exch<-matrix(0,nrow=4,ncol=2)
    ar<-qls(data1[[num]],family="gaussian",correlation="ar1")
    exch<-qls(data1[[num]],family="gaussian",correlation="exchangeable")
    beta.robust.var.ar[num,]<-beta.robust.sd(data1[[num]],alpha=ar$alpha,beta=ar$beta,
    correlation="ar1")
    beta.robust.var.exch[num,]<-beta.robust.sd(data1[[num]],alpha=exch$alpha,beta=exch$beta,
    correlation="exchangeable")
    beta.ar[num,]<-ar$beta
    beta.exch[num,]<-exch$beta
    for(n in 1:4){
      CI.ar[n,]<-beta.ar[num,][n]+c(-1,1)*qnorm(1-alpha[t]/2)*beta.robust.var.ar[num,][n]^0.5)
    for(p in 1:4){
      CI.exch[p,]<-beta.exch[num,][p]+c(-1,1)*qnorm(1-alpha[t]/2)*beta.robust.var.exch[num,][
      p]^0.5)
    for(a in 1:4){
      ifelse(CI.ar[a,][1]<b[a]& b[a]<CI.ar[a,][2],n.ar[a]<-n.ar[a]+1,n.ar[a])
    for(c in 1:4){
      ifelse(CI.exch[c,][1]<b[c]& b[c]<CI.exch[c,][2],n.exch[c]<-n.exch[c]+1,n.exch[c])
    }
  }
  coverage.ar[t,]<-n.ar/loop coverage.exch[t,]<-n.exch/loop }

### P-P plots for different models
library(MASS)
options(width=60)
def.par<-par(no.readonly=TRUE)
par(mfrow=c(1,2))
alpha<-seq(0.01,0.85,by=0.05)
plot(1-alpha,coverage.exch[,1],type="n",main="P-P plots for
different models(beta0)",xlab="1-alpha",ylab="Coverage
Percentage",ylim=c(0,1),xlim=c(0,1)) lines(x=0:1,y= 0:1, col =
"gray", lty=3)
points(1-alpha,coverage.exch[,1],cex=1.2,pch=21,col="red")
points(1-alpha,coverage.wcr[,1],cex=1.2,pch=20,col="green")
points(1-alpha,coverage.gee[,1],cex=1.2,pch=24,col="black")
legend(0,1,legend=c("CWGEE model", "WCR model", "GEE

```

```

model"),cex=1.2,pch=c(21,20,24),col=c("red","green","black"))

plot(1-alpha,coverage.exch[,2],type="n",main="P-P plots for
different models (beta1)",xlab="1-alpha",ylab="Coverage
Percentage",ylim=c(0,1),xlim=c(0,1)) lines(x=0:1,y= 0:1, col
="gray", lty=3)
points(1-alpha,coverage.exch[,2],cex=1.2,pch=21,col="red")
points(1-alpha,coverage.wcr[,2],cex=1.2,pch=20,col="green")
points(1-alpha,coverage.gee[,2],cex=1.2,pch=24,col="black")
legend(0,1,legend=c("CWGEE model", "WCR model", "GEE
model"),cex=1.2,pch=c(21,20,24),col=c("red","green","black"))
par(def.par)

options(width=60)
def.par<-par(no.readonly=TRUE)
par(mfrow=c(1,2))
plot(1-alpha,coverage.exch[,3],type="n",main="P-P plots for
different models (beta2)",xlab="1-alpha",ylab="Coverage
Percentage",ylim=c(0,1),xlim=c(0,1)) lines(x=0:1,y= 0:1, col =
"gray", lty=3)
points(1-alpha,coverage.exch[,3],cex=1.2,pch=21,col="red")
points(1-alpha,coverage.wcr[,3],cex=1.2,pch=20,col="green")
points(1-alpha,coverage.gee[,3],cex=1.2,pch=24,col="black")
legend(0,1,legend=c("CWGEE model", "WCR model", "GEE
model"),cex=1.2,pch=c(21,20,24),col=c("red","green","black"))

plot(1-alpha,coverage.exch[,4],type="n",main="P-P plots for
different models (beta3)",xlab="1-alpha",ylab="Coverage
Percentage",ylim=c(0,1),xlim=c(0,1))
lines(x=0:1,y= 0:1, col="gray", lty=3)
points(1-alpha,coverage.exch[,4],cex=1.2,pch=21,col="red")
points(1-alpha,coverage.wcr[,4],cex=1.2,pch=20,col="green")
points(1-alpha,coverage.gee[,4],cex=1.2,pch=24,col="black")
legend(0,1,legend=c("CWGEE model", "WCR model", "GEE
model"),cex=1.2,pch=c(21,20,24),col=c("red","green","black"))
par(def.par)

### P-P plots for CWGEE model (Exchangeable and AR-M)
library(MASS) options(width=60) def.par<-par(no.readonly=TRUE)
par(mfrow=c(1,2)) alpha<-seq(0.01,0.85,by=0.05)

```

```

plot(1-alpha,coverage.exch[,1],type="n",main="P-P plots for CWGEE
model (Exchangeable)",xlab="1-alpha",ylab="Coverage
Percentage",ylim=c(0,1),xlim=c(0,1))
lines(x=0:1,y= 0:1, col ="gray", lty=3)
points(1-alpha,coverage.exch[,1],cex=1.2,pch=1,col="blue")
points(1-alpha,coverage.exch[,2],cex=1.2,pch=2,col="red")
points(1-alpha,coverage.exch[,3],cex=1.2,pch=3,col="green")
points(1-alpha,coverage.exch[,4],cex=1.2,pch=4,col="black")
legend(0,1,legend=c("beta0","beta1","beta2","beta3"),cex=1.2,
pch=c(1,2,3,4),col=c("blue","red","green","black"))

```

```

plot(1-alpha,coverage.ar[,1],type="n",main="P-P plots for CWGEE
model (AR-M)",xlab="1-alpha",ylab="Coverage
Percentage",ylim=c(0,1),xlim=c(0,1))
lines(x=0:1,y= 0:1, col = "gray", lty=3)
points(1-alpha,coverage.ar[,1],cex=1.2,pch=1,col="blue")
points(1-alpha,coverage.ar[,2],cex=1.2,pch=2,col="red")
points(1-alpha,coverage.ar[,3],cex=1.2,pch=3,col="green")
points(1-alpha,coverage.ar[,4],cex=1.2,pch=4,col="black")
legend(0,1,legend=c("beta0", "beta1","beta2","beta3"),cex=1.2,
pch=c(1,2,3,4),col=c("blue","red","green","black")) par(def.par)

```

CURRICULUM VITAE

NAME: Ming Wang

ADDRESS: Department of Bioinformatics and Biostatistics
School of Public Health, University of Louisville
Louisville, KY 40292

E-MAIL: wm_pku@hotmail.com

EDUCATION:

B.S., Applied Mathematics
September 2002 - June 2006
Peking University, Beijing, China, 101100

HONORS AND AWARDS:

- the certification of Model Person for social practice 2003-2004
- superior learner in academic and practical training class 2003-2004
- Meritorious, Mathematical Modeling Contest 2005
- Research Assistantship at University of Louisville 2006
- SAS Certified Base Programmer 2007
- SAS Certified Advanced Programmer 2008

MEMBERSHIP:

- an assistant of the minister at academic and practical department 2003
- a minister of Mathematical Financial Association 2005
- Student Union Member at mathematical department 2005
- Vice-minister of Chinese Student Scholar Association in Louisville 2006
- Golden Key International Honor Society since 2006
- ASA membership since 2006

Activities:

- Organize social practice in summer and undertaking competitions 2004
- Internship in branch of Chinese Bank for two weeks 2005
- Organize the competition on analog stock market 2005
- Take part in activities in and out of campus as volunteer since 2006

RESEARCH

EXPERIENCE:

- Research experience during undergraduate study:
 1. Phylogeny Estimation of a Large Number of Sequences with High Level of Similarities
 2. An Application of Game Theory in Investment Decision
- Research projects during graduate study:
 1. Reliability of Wrist Specific Outcome Measures in Patients with Palmar Plate Fixation of Distal Radius Fractures
Ming Wang, Dr. McCabe. In progression
 2. Analysis of clustered longitudinal data
Ming Wang, Dr. Maiying Kong, Dr. Somnath Datta. Thesis research

INTEREST AND SKILLS:

- Proficient in C, Visual Basic, Latex
- Familiar with MATLAB, SAS, SPSS, R, SQL, DATABASE
- Sport: Pingpang, Swimming, Badminton
- Interest: Psychology, Art, History, Sociology