8-2010

Comparison of different methods for longitudinal data with missing observations.

Lin Sun
University of Louisville

COMPARISON OF DIFFERENT METHODS FOR LONGITUDINAL DATA WITH MISSING OBSERVATIONS

By

Lin Sun

B.S., Statistics, Tongji University, CHINA, 2008

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

Aug 2010

COMPARISON OF DIFFERENT METHODS FOR LONGITUDINAL DATA WITH MISSING OBSERVATIONS

By

Lin Sun

B.S., Statistics, Tongji University, CHINA, 2008

A Thesis Approved on

July 27, 2010

by the following Thesis Committee:

Thesis Chair: Maiying Kong, PhD

Guy Brock, PhD

Yong Li, PhD

ACKNOWLEDGEMENTS

First, I would like to thank my thesis advisor Dr. Maiying Kong for her insightful guidance and Dr. Guy Brock and Dr. Yong Li for serving on my thesis committee. I also would like to thank all the faculty members and students in the Department of Bioinformatics and Biostatistics at University of Louisville for their encouragement and friendship. I learned a lot form all the courses during the past two years. In addition, I would like to thank all my friends here at Louisville for their friendship and support. I would like to thank my parents, Jihai Sun and Shuqiu Ji, and all my families and friends in China for supporting me financially and emotionally, so that I am able to obtain the Master Degree in the United States. The time and experience here have been a great fortune for me and I will be always thankful to all those who helped and supported me.

ABSTRACT

COMPARISON OF DIFFERENT METHODS FOR LONGITUDINAL DATA WITH MISSING OBSERVATIONS

Lin Sun

July 27, 2010

Longitudinal studies occupy an important role in scientific researches and clinical trials. When taking the analysis of longitudinal data, investigators are often confronted with missing data which will produce potential biases, even in well-controlled condition.

In the literature, missing data could be classified as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Generalized estimating equations (GEE), Linear mixed effects model (LME) and Pattern-mixture effect model (PME) are the commonly used analysis methods for longitudinal data. In the current work, we carried out simulations on evaluating the performances of the different methods on analyzing longitudinal data. Based on our simulations, we conclude that when missing is MCAR, all the methods give valid estimation; when missing is MAR, GEE and PME give biased estimating results, while LME provides valid estimation. The choice of the patterns in PME may cause biased results; and when missing is MNAR, none of these models works very well, however, the selection of the patterns in PME may deserve further investigation.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

TABLE PAGE

CHAPTER I

INTRODUCTION

A Review of statistical methods for longitudinal data analysis

Longitudinal studies are now commonly used in biology, psychology, public health and clinical research [1]. For example, in a randomized clinical trial, investigators often collect prospective longitudinal data on one or more endpoints in response to a particular intervention relative to a control condition.

There are many advantages of longitudinal studies over cross-sectional studies [2]. First, in order to achieve the same statistical power, fewer subjects are needed in longitudinal studies. This is because the repeated measurements from a single subject provide more information than a single measurement of a single subject. Second, in a longitudinal study, each subject can serve as his/her own control. Generally, intra-subject variability is much less than inter-subject variability. Third, investigators are able to separate timing effects (changes over time within subjects) from cohort effects (differences between subjects at the baseline). Finally, longitudinal studies can provide information about individual change, which could not be provided by cross-sectional studies. However, longitudinal studies are also having their own challenges. One of those is the presence of missing data which is the focus of this thesis.

B Missing data in longitudinal analysis

When taking the analysis of longitudinal data, investigators are often confronted with missing data which will produce potential biases, even in well-controlled condition. For example, patients may drop out due to the result of the experiment.

There are several types of missing data, such as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). To make it clear, let us denote $Y_i = (Y_{i1}, \dots, Y_{iT})^T$ be the full-data response vector. Associated with Y_i is a $T \times P$ matrix of covariates $X_i = \left({X_{i1}}^T, \dots, {X_{iT}}^T\right)^T$. Partition Y_i into its observed and missing components, such that $Y_i = \left({Y_{i,obs}}^T, {Y_{i,mis}}^T\right)^T$. Throughout, we assume that X_i is fully observed, and Y_i has at least one observed component. We introduce $\mathrm{R_i} = (\mathrm{R_{i1}}, \dots, \mathrm{R_{iT}})^{\mathrm{T}}$ as indicator variables, where $\mathrm{R_{it}} = 1$ if $\mathrm{Y_{it}}$ is observed and $\mathrm{R_{it}} = 0$ if $\mathrm{Y_{it}}$ is missing.

Missing completely at random (MCAR)

MCAR is that the missing component does neither depend on observed components nor on the unobserved components. That is the missing data indicators R_i are independent of both ${Y_{i,obs}}^T$ and ${Y_{i,mis}}^T$.

A less stringent case of MCAR is what Little [3] refers to as covariate-dependent MCAR. Namely, given the covariates X_i, missingness R_i is independent of observed ${Y_{i,obs}}^T$ and unobserved dependent variables ${Y_{i,mis}}^T$. An example of covariate-dependent MCAR is when the number of follow-up visits differs by individual due to staggered entry and administrative censoring at a fixed calendar time.

Missing at random (MAR)

MAR is that the missingness may depend on X and ${Y_{i,obs}}^T$, but is independent of ${Y_{i,mis}}^T$. That is, given $\left(X, {Y_{i,obs}}^T\right)$, the missingness depends on ${Y_{i,mis}}^T$. For example, among participants with the same covariate profile, those who are observed to be sicker (via their values of ${Y_{i,obs}}^T$) are more likely to have missing values, so long as their missingness probability does not further depend on their missing responses.

Missing not at random (MNAR)

MNAR is that the missingness may depend on both ${Y_{i,obs}}^T$ and ${Y_{i,mis}}^T$.

C Statistical methods for longitudinal studies

There are several different general approaches to longitudinal data analysis [2]. One of the approaches is to reduce repeated measurements to a single summary measurement. Probably the earliest example is the t-test by Student [4]. The Analysis of Variance (ANOVA) for the repeated measurements, which assumes compound symmetry (constant variances and covariates over time) are also commonly used [5]. However, these approaches do not permit missing data or different measurement occasions for different subjects. Generalized mixed-effects regression models can be applied to numerous distributed outcomes, such as, normally distributed continuous as well as categorical outcomes. They are also quite robust to missing data. The disadvantage of the generalized mixed-effects regression models is more computationally complex than quasi-likelihood methods. Generalized Estimation Equations (GEE) models are often used as a general and computationally convenient alternative to mixed-effects regression models. However,

GEE models have some limitation to incomplete longitudinal data [5]. Alternatively, pattern mixed effects (PME) models are quite often used for missing data analyses.

In this thesis, we specifically focus on drop-out missing data in longitudinal studies. We will investigate the performance of GEE, LME, and PME models with simulations.

CHAPTER II

STATISTICAL METHODS FOR LONGITUDINAL DATA WITH MISSING OBSERVATIONS

A Generalized estimating equations (GEE)

GEE models were originally developed by Liang and Zeger [6] [7] [8]. GEE method is based on quasi-likelihood estimation, which is an extension of maximum likelihood method [9] [10]. In terms of missing data, it assumes that the missing data are MCAR. A basic premise of the GEE approach is that the regression parameters $\boldsymbol{\beta}$ but not the variance-covariance matrix of the repeated measures are the research interests. In longitudinal study, let $\boldsymbol{Y_i} = (Y_{i1}, \ldots, Y_{iT})^T$ be the response vector for subject i. Associated with $\boldsymbol{Y_i}$ is a $T \times P$ matrix of covariates $\boldsymbol{X_i} = \left({X_{i1}}^T, \ldots, {X_{iT}}^T\right)^T$, where $i = 1, 2, \cdots, N$. A generalized linear estimating equations could be described by $E(\boldsymbol{Y_i}) = \boldsymbol{\mu_i}$ and $g(\boldsymbol{\mu_i}) = \boldsymbol{X_i\beta}$, where g is a link function to connect $\boldsymbol{\mu_i}$ and $\boldsymbol{X_i\beta}$.

When g is an identity function and $\boldsymbol{Y_i}$ follows normal distribution, the generalized estimating equations has the form

$$\boldsymbol{Y_i} = \boldsymbol{X_i\beta} + \boldsymbol{\varepsilon_i},\ \ (i = 1, 2, \cdots, N). \tag{1}$$

The variance is generally described as a function of mean

$$Var(\boldsymbol{Y_i}) = \phi var(\boldsymbol{\mu_i}) \tag{2}$$

where $var(\boldsymbol{\mu_i})$ is a known variance function and ϕ is a scale parameter that may be known or estimated. In the current work, we consider the case that $\boldsymbol{Y_i}$ is normally distributed with an identity link function and $\phi = 1$.

In addition, one needs to specify the "working correlation structure" $\boldsymbol{R_i}$ for GEE models, where $\boldsymbol{R_i}$ is a $n_i \times n_i$ matrix for a given Y_i. We assume that $\boldsymbol{R_i}$ depends on variance-covariance parameters, denoted $\boldsymbol{\alpha}$. The usual working correlations considered are independence, exchangeable, AR (1), m-dependent, and unspecified [8]. Define $\boldsymbol{A_i}$ to be the $n_i \times n_i$ diagonal matrix with the ith diagonal element as $var(\mu_{ij}) = var(\varepsilon_{ij})$, then we will have:

$$Var(\boldsymbol{Y_i}) = \boldsymbol{A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}}}. \tag{3}$$

The quasi-likelihood estimates of the regression parameter can be obtained by the estimating equation:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu_i(\beta)}}{\partial \boldsymbol{\beta}} \{\boldsymbol{y_i} - \boldsymbol{\mu_i(\beta)}\} = \boldsymbol{0}. \tag{4}$$

B Linear mixed effects model (LME)

In practice, longitudinal data are often highly unbalanced due to missing and/ or that measurements are not taken at fixed time points. A two-stage analysis linear mixed effects model is often applied.

Let Y_{ij} denote the response of interest, for the jth response of the ith individual, measured at time t_{ij}, $i = 1, 2, \cdots, N$, $j = 1, 2, \cdots, n_i$, and $\boldsymbol{Y_i} = \left(Y_{i1}, \ldots, Y_{in_i}\right)^T$. The first stage of the two-stage approach assumes that $\boldsymbol{Y_i}$ satisfies the linear regression model

$$\boldsymbol{Y_i} = \boldsymbol{Z_i}\boldsymbol{\beta_i} + \boldsymbol{\varepsilon_i}, \tag{5}$$

where $\boldsymbol{Z_i}$ is a $n_i \times q$ matrix of known covariates, modeling how the response changes over time for the ith subject. $\boldsymbol{\beta_i}$ is a q-dimensional vector of unknown subject-specific regression coefficients, and $\boldsymbol{\varepsilon_i}$ is a vector of residual components $\varepsilon_{ij}, j = 1, 2, \cdots, n_i$. It is usually assumed that all $\boldsymbol{\varepsilon_i}$ are independent and normally distributed with mean vector zero, and covariance matrix $\sigma^2\boldsymbol{I_{n_i}}$, where $\boldsymbol{I_{n_i}}$ is the n_i-dimensional identity matrix. The subject-specific regression coefficients could be written as a function of population parameters $\boldsymbol{\beta}$ and a random effect $\boldsymbol{b_i}$ in the following form

$$\boldsymbol{\beta_i} = \boldsymbol{K_i}\boldsymbol{\beta} + \boldsymbol{b_i}. \tag{6}$$

The random effect $\boldsymbol{b_i}$ is generally assumed to be normally distributed with mean zero and variance D. Σ_b is used to explain the observed variability between the subjects, with respect to their subject-specific regression coefficients $\boldsymbol{\beta_i}$. $\boldsymbol{K_i}$ is a $q \times p$ matrix of known covariates, and $\boldsymbol{\beta}$ is a p-dimensional vector of unknown regression parameters. $\boldsymbol{b_i}$ $(i = 1,2, \ldots N)$ are assumed to be independent.

In order to combine the models from the two-stage analysis, we replace $\boldsymbol{\beta_i}$ in (1) by expression (6), yielding

$$\boldsymbol{Y_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_ib_i} + \boldsymbol{\varepsilon_i}, \tag{7}$$

where $\boldsymbol{X_i} = \boldsymbol{Z_iK_i}$ is the appropriate $n_i \times p$ matrix of known covariates, and where all other components are as defined earlier. (7) is called liner mixed-effects model with fixed effects $\boldsymbol{\beta}$ and with subject-specific effects $\boldsymbol{b_i}$.

C Pattern-mixture effect model (PME)

Pattern-mixture models are frequently used for longitudinal data analysis with dropouts. These models stratify the data according to time to dropout and formulate a model for each stratum. For pattern-mixture models, model for the distribution of the outcomes given drop-out pattern is specified, and then combined with a model for dropout. It may be the case that the distribution of outcomes given drop-out pattern is not completely identifiable, since for some drop-out patterns certain outcomes are not observed.

The first step is to divide the subjects into groups depending on their missing-data pattern. For example, suppose that subjects are measured at five timepoints with none missing at the first timepoint, and we only consider monotone drop-out, then we have five possible missing-data patterns.

pattern group	Time 1	Time 2	Time 3	Time 4	Time 5
1	O	O	O	O	O
2	O	O	O	O	M
3	O	O	O	M	M
4	O	O	M	M	M
5	O	M	M	M	M

Figure 1. Missing-data pattern

In the simulation, we combine the last three pattern groups into one new pattern group and remain the first two pattern groups. We will have three new pattern groups. We apply LME to different pattern groups separately. We get $\boldsymbol{\beta}^{(1)}$, $\boldsymbol{\beta}^{(2)}$, $\boldsymbol{\beta}^{(3)}$, $\sigma^{2(1)}$, $\sigma^{2(2)}$ and $\sigma^{2(3)}$.

The final estimate is

$$\boldsymbol{\beta} = \frac{N_1\boldsymbol{\beta}^{(1)}+N_2\boldsymbol{\beta}^{(2)}+N_3\boldsymbol{\beta}^{(3)}}{N} \tag{9}$$

and

$$\sigma^2 = \frac{N_1\sigma^{2(1)}+N_2\sigma^{2(2)}+N_3\sigma^{2(3)}}{N}, \tag{10}$$

where N_i is number of subjects for the ith pattern, i=1, 2, 3. $N = N_1 + N_2 + N_3$.

CHAPTER III

SIMULATIONS AND RESULTS

A Simulation scenarios

We simulated data according to the following model

$$y_{ij} = \beta_0 + \beta_1 Time_j + \beta_2 Grp_i + \beta_3 (Grp_i \times Time_j) + v_{0i} + v_{1i} Time_j + \varepsilon_{ij}, \qquad (11)$$

where $Time_j$ was coded 0, 1, 2, 3, 4 for five timepoints, and Grp_i was a dummy-coded (i.e. , 0 or 1) grouping variable. The regression coefficients were defined as: $\beta_0 = 25$, $\beta_1 = -1$, $\beta_2 = 0$, and $\beta_3 = -1$. The random subject effects v_{0i} and v_{1i} were assumed to be normal with zero means, variances $\sigma^2_{v_0} = 4$ and $\sigma^2_{v_1} = 0.25$, and covariance $\sigma_{v_{01}} = -0.1$. ε_is were assumed to be normal with mean 0 and variance $\sigma^2 = 4$.

In the simulations, 500 subjects were generated, each with 5 timepoints according to varieties of missing mechanism described below. Three statistical models (i.e. , GEE, LME and PME) were applied to the resulting incomplete dataset. GEE model with an independence working correlation structure was applied.

For the MCAR situation, we simulated data with dropout rates of 0%, 25%, 50%, 75% and 87.5% for the five respective timepoints. If subject was missing at a given

timepoint, then it was missing at all later timepoints as well. These rates indicate the percentage of the original sample that were missing at each timepoint. The results are reported in Table 1.

For the covariate-dependent MCAR, we simulated a case where the MCAR specification was different for the two groups. Specially, a subject with group 0 dropped out at 2nd, 3rd, 4th, and 5th timepoint if $v_{1i} > 0.6407758$, $v_{1i} > 0.4208106$, $v_{1i} > 0.2622003$ and $v_{1i} > 0$, respectively. For group1, the subject dropped out at 2nd, 3rd, 4th, and 5th timepoint if $v_{1i} < -0.6407758$, $v_{1i} < -0.4208106$, $v_{1i} < -0.2622003$ and $v_{1i} < 0$, respectively. The results are reported in Table 2.

For MAR, if the value of the dependent variable was lower than 23, then the subject dropped out at the next timepoint. The results are reported in Table 3.

For MNAR, after the first timepoint, if the value of the dependent variable was lower than 21.5, the subject was missing at that timepoint and all subsequent timepoints. The results are reported in Table 4.

B Simulation results

TABLE 1

Simulation results for longitudinal data with MCAR

models	β_0	β_1	β_2	β_3	σ^2
GEE	**25.00203**	**-1.00094**	**0.003150**	**-1.001388**	
variance	0.033988	0.011308	0.059374	0.021624	
MSE	0.033925	0.011286	0.059265	0.021583	
LME	**25.00289**	**-1.00554**	**0.001401**	**-0.99746**	**2.001874**
variance	0.033374	0.007938	0.05826	0.01554	0.003426
MSE	0.033315	0.007953	0.058145	0.015515	0.003422
PME	**25.00246**	**-1.00427**	**0.001988**	**-0.99873**	**1.98619**
variance	0.034336	0.015789	0.061503	0.028588	0.005061
MSE	0.034273	0.015776	0.061384	0.028533	0.005242
True value	**25**	**-1**	**0**	**-1**	**2**

Notes: **GEE** indicates Generalized Estimating Equations; **LME** indicates Linear Mixed Effected Model; **PME** indicates Pattern-Mixture Effect Model; **variance**=$\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta})^2$; **MSE** is Mean Square Error, which is $\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_{true})^2$.

Based on the results in Table 1, when the missing is MCAR, the averages of the estimates from all the three methods (GEE, LME, and PME) are close to the underlying values, indicating the estimates are unbiased. From Table 1, the variances are close to the mean squares for error (MSE), indicating the variance estimates are valid.

TABLE 2

Simulation results for longitudinal data with covariate-dependent MCAR

models	β_0	β_1	β_2	β_3	σ^2
GEE	**25.16352**	**-1.31524**	**-0.32275**	**-0.36851**	
variance	0.027198	0.003817	0.063235	0.007357	
MSE	0.053884	0.103188	0.167275	0.406124	
LME	**25.1147098**	**-1.26932**	**-0.22332**	**-0.46143**	**2.007398**
variance	0.02701045	0.003185	0.063127	0.006673	0.001811
MSE	0.04011477	0.07571	0.112875	0.296718	0.001862
PME	**25.00732**	**-1.05878**	**-0.00794**	**-0.88578**	**1.977186**
variance	0.02775047	0.007338	0.063855	0.014365	0.002587
MSE	0.02774855	0.010778	0.06379	0.027381	0.003102
True value	**25**	**-1**	**0**	**-1**	**2**

Notes: **GEE** indicates Generalized Estimating Equations; **LME** indicates Linear Mixed Effected Model; **PME** indicates Pattern-Mixture Effect Model; **variance**=$\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta})^2$; **MSE** is Mean Square Error, which is $\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_{true})^2$.

Based on the results in Table 2, when the missing is covariate-dependent MCAR, the averages of the estimates from GEE and LME are a little bit different from the underlying values, and the variances are smaller than the mean squares for error (MSE), indicating the estimates from GEE and LME are unbiased. However, From Table 2, the estimates from PME are unbiased, and the estimates from PME seem better than those obtained from GEE and LME.

TABLE 3

Simulation results for longitudinal data with MAR

models	β_0	β_1	β_2	β_3	σ^2
GEE	**24.94087**	**-0.342307**	**-0.069148**	**-0.796353**	
variance	0.025098	0.006572	0.057791	0.019503	
MSE	0.028545	0.439118	0.062457	0.060936	
LME	**25.00252**	**-1.005601**	**0.00537**	**-1.004257**	**1.994399**
variance	0.027691	0.011293	0.06574	0.016706	0.004038
MSE	0.027642	0.011302	0.065637	0.016691	0.004061
PME	**24.85061**	**-1.456717**	**0.578533**	**-0.603406**	**1.884995**
variance	0.027416	0.016019	0.054481	0.019467	0.004159
MSE	0.049678	0.224578	0.389073	0.176715	0.017377
True value	**25**	**-1**	**0**	**-1**	**2**

Notes: **GEE** indicates Generalized Estimating Equations; **LME** indicates Linear Mixed Effected Model; **PME** indicates Pattern-Mixture Effect Model; **variance**=$\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta})^2$; **MSE** is Mean Square Error, which is $\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_{true})^2$.

Based on the results in Table 3, when the missing is MAR, the averages of the estimates from GEE and PME are a little bit different from the underlying values. However, the averages of the estimates from LME are close to the underlying values, indicating the estimates are unbiased. The results may be impacted by the choice of the patterns, which will be further investigated in our future study.

TABLE 4

Simulation results for longitudinal data with MNAR

models	β_0	β_1	β_2	β_3	σ^2
GEE	**25.03088**	**-0.18259**	**-0.07786**	**-0.33949**	
variance	0.023686	0.004078	0.050935	0.011756	
MSE	0.024592	0.672226	0.056895	0.448002	
LME	**24.97535**	**-0.26055**	**-0.04812**	**-0.45267**	**1.740934**
variance	0.023315	0.004022	0.051795	0.011359	0.002441
MSE	0.023876	0.550806	0.054006	0.310902	0.069551
PME	**24.66288**	**-0.28371**	**0.687482**	**-0.52205**	**1.654873**
variance	0.025556	0.012373	0.058426	0.021881	0.002942
MSE	0.139152	0.52542	0.530942	0.25027	0.122048
True value	**25**	**-1**	**0**	**-1**	**2**

Notes: **GEE** indicates Generalized Estimating Equations; **LME** indicates Linear Mixed Effected Model; **PME** indicates Pattern-Mixture Effect Model; **variance**=$\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta})^2$; **MSE** is Mean Square Error, which is $\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_{true})^2$.

Based on the results in Table 4, when the missing is MNAR, the averages of the estimates from all three methods are a little bit different from the underlying values, and the variances are smaller than the mean squares for error (MSE), indicating the estimates from GEE and LME are unbiased. However, the results may be impacted by the choice of the patterns, which will be further investigated in our future study.

CHAPTER IV

FUTURE WORK

According to the simulation results, we can choose GEE for MCAR and LME for MAR; However, for MNAR, neither GEE nor LME works very well. PME may work well, but it deserves further discussion in the choices of combinations of patterns. We will discuss this in our further work. Also we hope to find a better model dealing with MNAR.

In the practice, it is very difficult to tell the missing data mechanism of a set of real data, especially for the MNAR. In the future work, we will focus on this problem.

REFERENCES

[1] J. D. Singer and J. B. Willett, *Applied Longitudinal Data Analysis: Modeling Change And Event Occurrence*, Oxford University Press, 2002.

[2] D. Hedeker and R. D. Gibbons, *Longitudinal Data Analysis*, Wiley Publications, 2006

[3] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Wiley Publications, 2002

[4] W. S. Gosset, "The probable error of a mean." , *Biometrika*, vol. 6, pp. 1–25. 1908.

[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] K. Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models." , *Biometrika*, vol. 73, pp. 13-22, 1986.

[7] S. L. Zeger and K. Y. Liang, "Longitudinal data analysis for discrete and continuous outcomes." , *Biometrics*, vol. 42, pp. 121-130, 1986.

[8] S. L. Zeger, K. Y. Liang and P. S. Albert, "Models for longitudinal data: a generalized estimating equation approach." , *Biometrics*, vol. 44, pp. 1049-1060, 1988.

[9] R. W. M. Wedderburn, "Quasi-likelihood functions, generalized linear models, and the gauss-newton method." *Biometrika*, vol. 61, pp. 439-447, 1974.

[10] P. McCullagh and J. A. Nelder, "Quasi-likelihood functions." , *Annal of statistics*, vol. 11, pp. 59-67, 1983.

Appendix: R code for the thesis

```
##############################################################################################
##                                  Longitudinal data with missing data                       ##
##############################################################################################
install.packages("MASS")
install.packages("geepack")
install.packages("nlme")
library(MASS)
library(geepack)
library(nlme)

loop<-500
gee.beta.est<-lme.beta.est<-pat.beta.est <-matrix(0,loop,4)
lme.sigma.est<-pat.sigma.est <-c(1:loop)
beta<-c(25, -1, 0, -1)
var.v<-matrix(c(4, -0.1, -0.1, 0.25), nrow=2)
sigma<-2
t<-c(0,1,2,3,4)
y<-rep(0,2500)
index1<-index2<-index3<-index4<-c()
b1<-c(25,-1,0,-1,2)
b2<-c(25,-1,0,-1)
gee.mse<-c(0,0,0,0)
lme.mse<-pat.mse <-c(0,0,0,0,0)
##############################################################################################
##                                          MCAR                                            ##
##############################################################################################
set.seed(100)
for (num in 1:loop)
```

```
{
print(num)
 for(i in 1:500)
 {  if (i<=250) grp<-0
    if (i>250) grp<-1
    vi<-mvrnorm(n=1,mu=c(0,0),Sigma=var.v)
    err<-rnorm(5, 0, 2)
    temp<-beta[1]+beta[2]*t+beta[3]*grp+beta[4]*grp*t+
          vi[1]+vi[2]*t+err
    index1<-rbinom(1,1,0.75)
    index2<-rbinom(1,1,0.67)
    index3<-rbinom(1,1,0.5)
    index4<-rbinom(1,1,0.5)
    if (index1==0) temp[2:5]<-NA
    if (index1==1 && index2==0) temp[3:5]<-NA
    if (index1==1 && index2==1 && index3==0) temp[4:5]<-NA
    if (index1==1 && index2==1 && index3==1 && index4==0) temp[5]<-NA
    y[(5*(i-1)+1):(5*i)]<-temp
}

fulldata<-data.frame(y=y,ID=rep(1:500,each=5),
      time=rep(0:4,500),group=rep(0:1,each=1250), d=matrix(data = 0, nrow = 500, ncol = 4))
y.new<-fulldata$y
ind<-!is.na(y.new)
mcar<-fulldata[ind,]
data1<-groupedData(y~time*group|ID,data=mcar)

#####lme#####
lme<-
lme(fixed=y~time*group,data=data1,random=list(ID=pdSymm(~time)),control=lmeControl(returnObject=
TRUE))
```

```
lme.beta.est[num,]<-lme$coefficients$fixed

lme.sigma.est[num]<-lme$sigma


######gee1######

gee1 <- geeglm(y~time*group,data=data1,id=ID,

  family=gaussian("identity"),corstr="independence")

gee.beta.est[num,]<-gee1$coefficients


######pattern######

for (ID in 1:500){

ind<-mcar$ID==ID

if (max(mcar$time[ind])==4) {mcar$pattern[ind]<-3}

if (max(mcar$time[ind])==3) {mcar$pattern[ind]<-2}

if (max(mcar$time[ind])<=2) {mcar$pattern[ind]<-1}

            }

mcar<-groupedData(y~time*group|ID,data=mcar)


lme.pattern3<-
lme(fixed=y~time*group,data=mcar[mcar$pattern==3,],random=list(ID=pdSymm(~time)),control=lmeCon
trol(returnObject=TRUE))

n<-length(unique(mcar$ID))

n3<-length(unique(mcar$ID[mcar$pattern==3]))


lme.pattern2<-
lme(fixed=y~time*group,data=mcar[mcar$pattern==2,],random=list(ID=pdSymm(~time)),control=lmeCon
trol(returnObject=TRUE))

n2<-length(unique(mcar$ID[mcar$pattern==2]))


lme.pattern1<-
lme(fixed=y~time*group,data=mcar[mcar$pattern==1,],random=list(ID=pdSymm(~time)),control=lmeCon
trol(returnObject=TRUE))

n1<-length(unique(mcar$ID[mcar$pattern==1]))
```

```
temp.beta.est3<-lme.pattern3$coefficients$fixed
temp.beta.est2<-lme.pattern2$coefficients$fixed
temp.beta.est1<-lme.pattern1$coefficients$fixed
temp.beta.est<-(n1*temp.beta.est1+n2*temp.beta.est2+n3*temp.beta.est3)/n
temp.sigma.est3<-lme.pattern3$sigma
temp.sigma.est2<-lme.pattern2$sigma
temp.sigma.est1<-lme.pattern1$sigma
temp.sigma.est<-(n1*temp.sigma.est1+n2*temp.sigma.est2+n3*temp.sigma.est3)/n
pat.beta.est[num,]<-temp.beta.est
pat.sigma.est[num]<-temp.sigma.est
}
#####result######
gee.result<-gee.beta.est
lme.result<-cbind(I=lme.beta.est,sigma=lme.sigma.est)
pat.result<-cbind(I=pat.beta.est,sigma=pat.sigma.est)

######mean####
gee.mean<-apply(gee.result,2,mean)
lme.mean<-apply(lme.result,2,mean)
pat.mean<-apply(pat.result,2,mean)

######var####
gee.var<-apply(gee.result,2,var)
lme.var<-apply(lme.result,2,var)
pat.var<-apply(pat.result,2,var)

######mse####
for( j in 1:4){
gee.mse[j]<-sum((gee.result[,j]-b2[j])^2)/loop
        }
```

```
for( j in 1:5){
lme.mse[j]<-sum((lme.result[,j]-b1[j])^2)/loop
        }
for( j in 1:5){
pat.mse[j]<-sum((pat.result[,j]-b1[j])^2)/loop
        }

gee.final<-rbind(gee.mean=gee.mean,gee.var=gee.var,gee.mse=gee.mse)
lme.final<-rbind(lme.mean=lme.mean,lme.var=lme.var,lme.mse=lme.mse)
pat.final<-rbind(pat.mean=pat.mean,pat.var=pat.var,pat.mse=pat.mse)
gee.final
lme.final
pat.final
##############################################################################################
##                                covariate-dependent MCAR                                  ##
##############################################################################################
set.seed(999)
c<-
c(qnorm(0.1,mean=0,sd=0.5),qnorm(0.2,mean=0,sd=0.5),qnorm(0.3,mean=0,sd=0.5),qnorm(0.5,mean=0,s
d=0.5))
c1<-(-c)

for (num in 1:loop)
{
print(num)
 for(i in 1:500)
 {  if (i<=250) grp<-0
    if (i>250) grp<-1
    vi<-mvrnorm(n=1,mu=c(0,0),Sigma=var.v)
    err<-rnorm(5, 0, 2)
    temp<-beta[1]+beta[2]*t+beta[3]*grp+beta[4]*grp*t+
```

```
        vi[1]+vi[2]*t+err

    if (grp==0 && vi[2]>=c1[1]){temp[2:5]<-NA}
    if (grp==0 && vi[2]<c1[1] && vi[2]>=c1[2]){temp[3:5]<-NA}
    if (grp==0 && vi[2]<c1[2] && vi[2]>=c1[3]){temp[4:5]<-NA}
    if (grp==0 && vi[2]<c1[3] && vi[2]>=c1[4]){temp[5]<-NA}
    if (grp==1 && vi[2]<=c[1]){temp[2:5]<-NA}
    if (grp==1 && vi[2]>c[1] && vi[2]<=c[2]){temp[3:5]<-NA}
    if (grp==1 && vi[2]>c[2] && vi[2]<=c[3]){temp[4:5]<-NA}
    if (grp==1 && vi[2]>c[3] && vi[2]<=c[4]){temp[5]<-NA}
    y[(5*(i-1)+1):(5*i)]<-temp
 }

fulldata<-data.frame(y=y,ID=rep(1:500,each=5),
     time=rep(0:4,500),group=rep(0:1,each=1250), d=matrix(data = 0, nrow = 500, ncol = 4))
y.new<-fulldata$y
ind<-!is.na(y.new)
mnar<-fulldata[ind,]
data1<-groupedData(y~time*group|ID,data=mnar)

#####lme#####
lme<-
lme(fixed=y~time*group,data=data1,random=list(ID=pdSymm(~time)),control=lmeControl(returnObject=
TRUE))
lme.beta.est[num,]<-lme$coefficients$fixed
lme.sigma.est[num]<-lme$sigma

######gee1######
gee1 <- geeglm(y~time*group,data=data1,id=ID,
  family=gaussian("identity"),corstr="independence")
gee.beta.est[num,]<-gee1$coefficients
```

```
######pattern######

for (ID in 1:500){

ind<-mnar$ID==ID

if (max(mnar$time[ind])==4) {mnar$pattern[ind]<-3}

if (max(mnar$time[ind])==3) {mnar$pattern[ind]<-2}

if (max(mnar$time[ind])<=2) {mnar$pattern[ind]<-1}

        }

mnar<-groupedData(y~time*group|ID,data=mnar)


lme.pattern3<-
lme(fixed=y~time*group,data=mnar[mnar$pattern==3,],random=list(ID=pdSymm(~time)),control=lmeCo
ntrol(returnObject=TRUE))

n<-length(unique(mnar$ID))

n3<-length(unique(mnar$ID[mnar$pattern==3]))


lme.pattern2<-
lme(fixed=y~time*group,data=mnar[mnar$pattern==2,],random=list(ID=pdSymm(~time)),control=lmeCo
ntrol(returnObject=TRUE))

n2<-length(unique(mnar$ID[mnar$pattern==2]))


lme.pattern1<-
lme(fixed=y~time*group,data=mnar[mnar$pattern==1,],random=list(ID=pdSymm(~time)),control=lmeCo
ntrol(returnObject=TRUE))

n1<-length(unique(mnar$ID[mnar$pattern==1]))


temp.beta.est3<-lme.pattern3$coefficients$fixed

temp.beta.est2<-lme.pattern2$coefficients$fixed

temp.beta.est1<-lme.pattern1$coefficients$fixed

temp.beta.est<-(n1*temp.beta.est1+n2*temp.beta.est2+n3*temp.beta.est3)/n

temp.sigma.est3<-lme.pattern3$sigma

temp.sigma.est2<-lme.pattern2$sigma

temp.sigma.est1<-lme.pattern1$sigma
```

```
temp.sigma.est<-(n1*temp.sigma.est1+n2*temp.sigma.est2+n3*temp.sigma.est3)/n
pat.beta.est[num,]<-temp.beta.est
pat.sigma.est[num]<-temp.sigma.est

#####result######
gee.result<-gee.beta.est
lme.result<-cbind(I=lme.beta.est,sigma=lme.sigma.est)
pat.result<-cbind(I=pat.beta.est,sigma=pat.sigma.est)

######mean####
gee.mean<-apply(gee.result,2,mean)
lme.mean<-apply(lme.result,2,mean)
pat.mean<-apply(pat.result,2,mean)

######var####
gee.var<-apply(gee.result,2,var)
lme.var<-apply(lme.result,2,var)
pat.var<-apply(pat.result,2,var)
###### mse######

for( j in 1:4){
gee.mse[j]<-sum((gee.result[,j]-b2[j])^2)/loop
        }
for( j in 1:5){
lme.mse[j]<-sum((lme.result[,j]-b1[j])^2)/loop
        }
for( j in 1:5){
pat.mse[j]<-sum((pat.result[,j]-b1[j])^2)/loop
        }
```

```
gee.final<-rbind(gee.mean=gee.mean,gee.var=gee.var,gee.mse=gee.mse)
lme.final<-rbind(lme.mean=lme.mean,lme.var=lme.var,lme.mse=lme.mse)
pat.final<-rbind(pat.mean=pat.mean,pat.var=pat.var,pat.mse=pat.mse)

gee.final
lme.final
pat.final
##############################################################################################
##                                          MAR                                             ##
##############################################################################################
set.seed(999)
for (num in 1:loop)
{
print(num)
 for(i in 1:500)
 {  if (i<=250) grp<-0
    if (i>250) grp<-1
    vi<-mvrnorm(n=1,mu=c(0,0),Sigma=var.v)
    err<-rnorm(5, 0, 2)
    temp<-beta[1]+beta[2]*t+beta[3]*grp+beta[4]*grp*t+
          vi[1]+vi[2]*t+err

      if (temp[1]<23)  temp[2:5]<-NA
      else {if (temp[2]<23)  temp[3:5]<-NA
      else {if (temp[3]<23)  temp[4:5]<-NA
      else {if (temp[4]<23)  temp[5]<-NA }}}

    y[(5*(i-1)+1):(5*i)]<-temp
 }
```

```
fulldata<-data.frame(y=y,ID=rep(1:500,each=5),
      time=rep(0:4,500),group=rep(0:1,each=1250), d=matrix(data = 0, nrow = 500, ncol = 4))
y.new<-fulldata$y
ind<-!is.na(y.new)
mar<-fulldata[ind,]
data1<-groupedData(y~time*group|ID,data=mar)


######gee1######
gee1 <- geeglm(y~time*group,data=data1,id=ID,
  family=gaussian("identity"),corstr="independence")
gee.beta.est[num,]<-gee1$coefficients



#####lme#####
lme<-
lme(fixed=y~time*group,data=data1,random=list(ID=pdSymm(~time)),control=lmeControl(returnObject=
TRUE))


lme.beta.est[num,]<-lme$coefficients$fixed
lme.sigma.est[num]<-lme$sigma


######pattern######
for (ID in 1:500){
ind<-mar$ID==ID
if (max(mar$time[ind])==4) {mar$pattern[ind]<-3}
if (max(mar$time[ind])==3) {mar$pattern[ind]<-2}
if (max(mar$time[ind])<=2) {mar$pattern[ind]<-1}
          }
mar<-groupedData(y~time*group|ID,data=mar)
```

```
lme.pattern3<-
lme(fixed=y~time*group,data=mar[mar$pattern==3,],random=list(ID=pdSymm(~time)),control=lmeContr
ol(returnObject=TRUE))

n<-length(unique(mar$ID))

n3<-length(unique(mar$ID[mar$pattern==3]))


lme.pattern2<-
lme(fixed=y~time*group,data=mar[mar$pattern==2,],random=list(ID=pdSymm(~time)),control=lmeContr
ol(returnObject=TRUE))

n2<-length(unique(mar$ID[mar$pattern==2]))


lme.pattern1<-
lme(fixed=y~time*group,data=mar[mar$pattern==1,],random=list(ID=pdSymm(~time)),control=lmeContr
ol(returnObject=TRUE))

n1<-length(unique(mar$ID[mar$pattern==1]))


temp.beta.est3<-lme.pattern3$coefficients$fixed

temp.beta.est2<-lme.pattern2$coefficients$fixed

temp.beta.est1<-lme.pattern1$coefficients$fixed

temp.beta.est<-(n1*temp.beta.est1+n2*temp.beta.est2+n3*temp.beta.est3)/n

temp.sigma.est3<-lme.pattern3$sigma

temp.sigma.est2<-lme.pattern2$sigma

temp.sigma.est1<-lme.pattern1$sigma

temp.sigma.est<-(n1*temp.sigma.est1+n2*temp.sigma.est2+n3*temp.sigma.est3)/n

pat.beta.est[num,]<-temp.beta.est

pat.sigma.est[num]<-temp.sigma.est

}


#####result######

gee.result<-gee.beta.est

lme.result<-cbind(I=lme.beta.est,sigma=lme.sigma.est)

pat.result<-cbind(I=pat.beta.est,sigma=pat.sigma.est)
```

```
######mean####
gee.mean<-apply(gee.result,2,mean)
lme.mean<-apply(lme.result,2,mean)
pat.mean<-apply(pat.result,2,mean)

######var####
gee.var<-apply(gee.result,2,var)
lme.var<-apply(lme.result,2,var)
pat.var<-apply(pat.result,2,var)

##### mse######
for( j in 1:4){
gee.mse[j]<-sum((gee.result[,j]-b2[j])^2)/loop
        }
for( j in 1:5){
lme.mse[j]<-sum((lme.result[,j]-b1[j])^2)/loop
        }
for( j in 1:5){
pat.mse[j]<-sum((pat.result[,j]-b1[j])^2)/loop
        }

gee.final<-rbind(gee.mean=gee.mean,gee.var=gee.var,gee.mse=gee.mse)
lme.final<-rbind(lme.mean=lme.mean,lme.var=lme.var,lme.mse=lme.mse)
pat.final<-rbind(pat.mean=pat.mean,pat.var=pat.var,pat.mse=pat.mse)

gee.final
lme.final
pat.final
#############################################################################################
##                                        MNAR                                             ##
```

```
##############################################################################################
set.seed(500)
for (num in 1:loop)
{
print(num)
 for(i in 1:500)
 {  if (i<=250) grp<-0
    if (i>250) grp<-1
    vi<-mvrnorm(n=1,mu=c(0,0),Sigma=var.v)
    err<-rnorm(5, 0, 2)
    temp<-beta[1]+beta[2]*t+beta[3]*grp+beta[4]*grp*t+
          vi[1]+vi[2]*t+err

      if (temp[2]<21.5)  temp[2:5]<-NA
      else {if (temp[3]<21.5)  temp[3:5]<-NA
      else {if (temp[4]<21.5)  temp[4:5]<-NA
      else {if (temp[5]<21.5)  temp[5]<-NA}}}
   y[(5*(i-1)+1):(5*i)]<-temp
 }

fulldata<-data.frame(y=y,ID=rep(1:500,each=5),
     time=rep(0:4,500),group=rep(0:1,each=1250), d=matrix(data = 0, nrow = 500, ncol = 4))
y.new<-fulldata$y
ind<-!is.na(y.new)
mnar<-fulldata[ind,]
data1<-groupedData(y~time*group|ID,data=mnar)

#####lme#####
lme<-
lme(fixed=y~time*group,data=data1,random=list(ID=pdSymm(~time)),control=lmeControl(returnObject=
TRUE))
```

```
lme.beta.est[num,]<-lme$coefficients$fixed

lme.sigma.est[num]<-lme$sigma


######gee1######

gee1 <- geeglm(y~time*group,data=data1,id=ID,

  family=gaussian("identity"),corstr="independence")

gee.beta.est[num,]<-gee1$coefficients


######pattern######

for (ID in 1:500){

ind<-mnar$ID==ID

if (max(mnar$time[ind])==4) {mnar$pattern[ind]<-3}

if (max(mnar$time[ind])==3) {mnar$pattern[ind]<-2}

if (max(mnar$time[ind])<=2) {mnar$pattern[ind]<-1}

            }

mnar<-groupedData(y~time*group|ID,data=mnar)


lme.pattern3<-
lme(fixed=y~time*group,data=mnar[mnar$pattern==3,],random=list(ID=pdSymm(~time)),control=lmeCo
ntrol(returnObject=TRUE))

n<-length(unique(mnar$ID))

n3<-length(unique(mnar$ID[mnar$pattern==3]))


lme.pattern2<-
lme(fixed=y~time*group,data=mnar[mnar$pattern==2,],random=list(ID=pdSymm(~time)),control=lmeCo
ntrol(returnObject=TRUE))

n2<-length(unique(mnar$ID[mnar$pattern==2]))


lme.pattern1<-
lme(fixed=y~time*group,data=mnar[mnar$pattern==1,],random=list(ID=pdSymm(~time)),control=lmeCo
ntrol(returnObject=TRUE))

n1<-length(unique(mnar$ID[mnar$pattern==1]))
```

```
temp.beta.est3<-lme.pattern3$coefficients$fixed
temp.beta.est2<-lme.pattern2$coefficients$fixed
temp.beta.est1<-lme.pattern1$coefficients$fixed
temp.beta.est<-(n1*temp.beta.est1+n2*temp.beta.est2+n3*temp.beta.est3)/n
temp.sigma.est3<-lme.pattern3$sigma
temp.sigma.est2<-lme.pattern2$sigma
temp.sigma.est1<-lme.pattern1$sigma
temp.sigma.est<-(n1*temp.sigma.est1+n2*temp.sigma.est2+n3*temp.sigma.est3)/n
pat.beta.est[num,]<-temp.beta.est
pat.sigma.est[num]<-temp.sigma.est
}

#####result######
gee.result<-gee.beta.est
lme.result<-cbind(I=lme.beta.est,sigma=lme.sigma.est)
pat.result<-cbind(I=pat.beta.est,sigma=pat.sigma.est)

######mean####
gee.mean<-apply(gee.result,2,mean)
lme.mean<-apply(lme.result,2,mean)
pat.mean<-apply(pat.result,2,mean)

######var####
gee.var<-apply(gee.result,2,var)
lme.var<-apply(lme.result,2,var)
pat.var<-apply(pat.result,2,var)

######mse######
for( j in 1:4){
gee.mse[j]<-sum((gee.result[,j]-b2[j])^2)/loop
```

```
        }
for( j in 1:5){
lme.mse[j]<-sum((lme.result[,j]-b1[j])^2)/loop
        }
for( j in 1:5){
pat.mse[j]<-sum((pat.result[,j]-b1[j])^2)/loop
        }
gee.final<-rbind(gee.mean=gee.mean,gee.var=gee.var,gee.mse=gee.mse)
lme.final<-rbind(lme.mean=lme.mean,lme.var=lme.var,lme.mse=lme.mse)
pat.final<-rbind(pat.mean=pat.mean,pat.var=pat.var,pat.mse=pat.mse)

gee.final
lme.final
pat.final
################################################################################
```

CURRICULM VITAE

NAME: LIN SUN

ADDRESS:

Department of Bioinformatics and Biostatistics
School of Public Health and Information Science
University of Louisville
Louisville, KY 40292

E-MAIL: l0sun006@louiville.edu

EDUCATION:

B.S., Statistics
September 2004 – June 2008
Tongji University, Shanghai, CHINA, 200092

HONORS AND AWARDS:

- Tongji University Scholarship 2004-2006
- SAS certified Base Programmer Jun 2009

MEMBERSHIP:

- Student Union Member 2004-2006
- ASA Membership Since 2008

RESEARCH EXPERIENCE:

- Research experience during undergraduate study:

  1. Summer Internship in Local Statistical Bureau
  2. Statistical Analysis of Graduates' Employment at Tongji University

- Research projects and experience during graduate study

1. Safety and injury among teens enrolled in school-to-work apprentice programs
2. Sample Size Estimation in Longitudinal Data Analysis

INTEREST AND SKILLS:

- Proficiency with statistical software particularly SAS and R
- Working knowledge of office tools