Doubly robust generalized estimating equations with consistent variance estimator when one auxiliary model is misspecified

by

Dae-Gyu Jang

A Dissertation
submitted in fulfillment of the requirement
for the degree of
Master of Science
in
Statistics

The Department of Statistics
College of Natural Sciences
Seoul National University
August, 2016

Abstract

Dae-Gyu Jang
The Department of Statistics
The Graduate School
Seoul National University

The doubly-robust generalized estimating equation (DR-GEE) is a popular analytic tool for repeated measurements with missing data. It requires two assumptions on auxiliary models for outcome and the observation indicator, and produces a consistent point estimator when either of the model assumptions is correct. Another advantage is an easy variance formula due to asymptotic independence between the estimator of interest and the estimators from the auxiliary models. However, this feature can be capitalized only when both models are correctly specified. In this paper, we propose an alternative DR-GEE that produces a consistent variance estimator even when one of the auxiliary models is misspecified. The main idea is to construct estimating equations for the auxiliary models so that the estimators of main interest and the estimators from auxiliary models are asymptotically independent. We illustrate the method on data from a clinical trial for hypertension.

Keywords : Doubly robustness, Generalized estimating equations, Missing at random, Consistent variance estimator.

Student Number : 2014-20300

Contents

List of Tables

Chapter 1

Introduction

Generalized estimating equations (GEE) is widely used for inference in marginal models for longitudinal data analyses (Liang and Zeger, 1986). It is known that GEE method is valid under missing completely at random(MCAR) but is not generally valid under missing at random (MAR) mechanism. Under MAR, inverse probability weighting (Robins, Rotnitzky and Zhao, 1995) and imputation (Paik, 1997) have been developed as procedures that yield valid inference for GEE models. Correct dropout model is required for inverse probability weighting procedure, and a correct specification of the imputation model is required for the imputation procedure. The doubly robust generalized estimating equation (DR-GEE) is now a popular tool for analyzing repeated measures with missing data. The Imputation model and the missing probability model is specified for DR-GEE, and it produces a consistent point estimator when either auxiliary model is correct, known as the 'double robustness' property. The point estimate is also asymptotically optimal if both models are correctly specified. Another advantage of the DR-GEE is that asymptotic independence between the estimator of interest and the estimators from the auxiliary models result in an easy variance formula, as proposed by Seaman and Copas (2009). However, this feature can be capitalized only when both models are correctly specified. In this paper, we propose an alternative DR-GEE that produces a consistent variance estimator even when one of the auxiliary models is misspecified. This is done by constructing the estimating equations for the auxiliary models so that the estimators of main interest and the estimators from auxiliary models are asymptotically independent.

The motivating example is from a study of clinical trial in antihypertensive drug in isolated hypertension in the elderly. This is a randomized double blind placebo controlled study that was conducted across centers in the United States to assess the effect of antihypertensive drug treatment to reduce the risk of strokes. We are interested in investigating the treatment effect on mean arterial pressure across the three time points, where there are some patient dropouts in the followup period. We want to assume a missing at random mechanism and use the DR-GEE method to analyze this data with monotone missingness pattern. With the possibility that one of the auxiliary models are misspecified, the alternative DR-GEE still yields a consistent variance estimator when applied to this study. Details of the proposed method is presented in the following sections. We also include a simulations study to evaluate the method and apply it to the data from clinical trial in antihypertensive drug.

In chapter 2, we describe the settings more explicitly. In chapter 3, we propose our new DR-GEE. Then in chapter 4, we conduct a simulation study to compare bias and coverage probability of point estimator. It describes consistency of variance estimator of proposed method. In chapter 5, we apply the new method to the clinical trial in antihypertensive drug.

Chapter 2

Setup and Review

This chapter is for basic setup of our problem, and also a review of some existing methods. In 2.1, we describe some definitions and model assumptions. Then we explain GEE, and its modified versions, weighted GEE (WGEE), and DR-GEE. Finally, we explain the weak point of existing method, where our method is motivated.

2.1 Setup

We denote Y_{ij} to be the response of individual i at time j $(i = 1, \cdots, N$; $j = 1, \cdots, M)$, and $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{iM})^T$. $\boldsymbol{X}_{ij}$ is a covariate vector for individual i at time j, and $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}^T, \cdots, \boldsymbol{X}_{iM}^T)^T$, and $\boldsymbol{\beta}$ is the vector of parameters. We assume that covariates are fully observed. Let $\mu_{ij} :=\mathrm{E}(Y_{ij}|\boldsymbol{X}_{ij})$, and $g(\mu_{ij}) := \boldsymbol{X}_{ij}\boldsymbol{\beta}^T$ for some link function g, and let $\boldsymbol{\mu}_i = (\mu_{i1}, \cdots, \mu_{iM})^T$. We assume that $\boldsymbol{Y}_i \perp \boldsymbol{Y}_k | \boldsymbol{X}_i, \boldsymbol{X}_k$ for $i \neq k$.
In addition, we use the following notation for Weighted GEE: Let $\boldsymbol{L}_{ij}$ be a vector of the data collected on individual i at time j. $\bar{\boldsymbol{L}}_{ij} = (\boldsymbol{L}_{i1}^T, \cdots, \boldsymbol{L}_{ij}^T)^T$ is

the data available on individual i at time j, and $\underline{\boldsymbol{L}}_{ij} = (\boldsymbol{L}_{i,j+1}^T, \cdots, \boldsymbol{L}_{iM}^T)^T$ is the data not yet available on individual i at time j. n_i denotes the last observation time of individual i, and $C_{ij} = I(n_i = j)$; $\boldsymbol{C}_i = (C_{i1}, \cdots, C_{iM})^T$. R_{ij} is the response indicator where $R_{ij} = 1$ if Y_{ij} is observed, and 0 for missing. In this paper we assume that the pattern follows a monotone dropout, and that mechanism is MAR, so that $P\left(R_{ij} = 1|\bar{\boldsymbol{L}}_{iM}, R_{i,j-1} = 1\right) = P\left(R_{ij} = 1|\bar{\boldsymbol{L}}_{i,j-1}, R_{i,j-1} = 1\right)$.

2.2 Generalized Estimating Equations

The GEE for complete data are

$$\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} V_i(\boldsymbol{\rho})^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = 0 \tag{2.1}$$

where $V_i(\boldsymbol{\rho})$ is the working covariance matrix of $\boldsymbol{Y}_i$. When there are missing values in $\boldsymbol{Y}_i$, the standard GEE method uses eqations (2.1), with $\boldsymbol{Y}_i$, $\boldsymbol{\mu}_i$ and V_i replaced by their subvectors (submatrix) with respect to the observed data respectively. Under suitable regularity conditions, these observed-data GEE yield consistent estimator of $\boldsymbol{\beta}$ when the data are missing completely at random (MCAR) and $E(Y_{ij}|\boldsymbol{X}_i) = E(Y_{ij}|\boldsymbol{X}_{ij})\ \forall j$.

2.3 Weighted GEE

The WGEE are

$$\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} V_i(\boldsymbol{\rho})^{-1}\Delta(\boldsymbol{\alpha})(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = 0, \tag{2.2}$$

where $\Delta(\boldsymbol{\alpha}) = diag(R_{i1}/\pi_{i1}, \cdots, R_{iM}/\pi_{iM})$ and $\pi_{ij} = \pi_{ij}(\boldsymbol{\alpha}) = P(R_{ij} = 1 \mid \bar{\boldsymbol{L}}_{i,j-1})$ is the probability that $\boldsymbol{Y}_{ij}$ is observed. $\pi_{ij} = \prod_{k=1}^{j} p_{ik}$, where

$p_{ij} = P(R_{ij} = 1 \mid \bar{\boldsymbol{L}}_{i,j-1}, R_{i,j-1} = 1)$ is the probability that $\boldsymbol{Y}_{ij}$ is observed conditioned that $\boldsymbol{Y}_{i,j-1}$ is observed. Under suitable regularity conditions, the solution of (2.2) is consistent estimator of $\boldsymbol{\beta}$ when the data are MAR and the drop out model is correctly specified.

2.4 DR-GEE

The DR-GEE are

$$\sum_{i=1}^{N}\left\{\frac{R_{iM}}{\pi_{iM}}\frac{\partial\boldsymbol{\mu}_i^T}{\partial\boldsymbol{\beta}}V_i(\boldsymbol{\rho})^{-1}(\boldsymbol{Y}_i-\boldsymbol{\mu}_i)+\sum_{j=1}^{M-1}\frac{C_{ij}-\lambda_{i,j+1}R_{ij}}{\pi_{i,j+1}}\frac{\partial\boldsymbol{\mu}_i^T}{\partial\boldsymbol{\beta}}V_i(\boldsymbol{\rho})^{-1}(\boldsymbol{M}_{ij}-\boldsymbol{\mu}_i)\right\}=0, \tag{2.3}$$

where $M_{ij} = M_{ij}(\boldsymbol{\gamma}) = E\left[\boldsymbol{Y}_i \mid \bar{\boldsymbol{L}}_{ij},\ R_{ij} = 1\right]$ is subject to imputation model, $\lambda_{ij} = \lambda_{ij}(\boldsymbol{\phi}) = P(R_{ij} = 0 \mid \bar{\boldsymbol{L}}_{i,j-1},\ R_{i,j-1} = 1)$ and $\pi_{ij} = \pi_{ij}(\boldsymbol{\phi}) = P(R_{ij} = 1 \mid \bar{\boldsymbol{L}}_{i,j-1}) = \prod_{k=1}^{j}(1-\lambda_{ik})$. Under suitable regularity conditions and MAR condition, the solution of (2.3) is consistent estimator of $\boldsymbol{\beta}$ when either the dropout model or the imputation model is correctly specified (Doubly Robustness). Moreover, when both the auxiliary models are correctly specfied, the solution of (2.3) is asymptotically optimal.

The variance formula for DR GEE proposed by Seaman and Copas (2009) is

$$\widehat{var(\hat{\boldsymbol{\beta}})} = \left\{\sum_{i=1}^{N}A_i(\hat{\boldsymbol{\beta}})\right\}^{-1}\left\{\sum_{i=1}^{N}U_i(\hat{\boldsymbol{\beta}})U_i(\hat{\boldsymbol{\beta}})^T\right\}\left\{\sum_{i=1}^{N}A_i(\hat{\boldsymbol{\beta}})\right\}^{-1} \tag{2.4}$$

where

$$\begin{aligned}
U_i(\boldsymbol{\beta}) &= \frac{R_{iM}}{\pi_{iM}}\frac{\partial\boldsymbol{\mu}_i^T}{\partial\boldsymbol{\beta}}V_i(\boldsymbol{\rho})^{-1}(\boldsymbol{Y}_i-\boldsymbol{\mu}_i)\\
&\quad+\sum_{j=1}^{M-1}\frac{C_{ij}-\lambda_{i,j+1}R_{ij}}{\pi_{i,j+1}}V_i(\boldsymbol{\rho})^{-1}(\boldsymbol{M}_{ij}-\boldsymbol{\mu}_i),\\
A_i(\boldsymbol{\beta}) &= E\left[-\frac{\partial U_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}\right] = \frac{\partial\boldsymbol{\mu}_i^T}{\partial\boldsymbol{\beta}}V_i(\boldsymbol{\rho})^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}^T}.
\end{aligned}$$

In the original DR-GEE, the variance formula (2.4) is consistent when both the auxiliary models are correctly specified. However, when one of the auxiliary models is mis-specified, consistency of (2.4) is ungarenteed.

In the original DR-GEE, there is no guide line for estimating equations for $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$. Let $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\phi}}$ be $\sqrt{N}-$consistent estimator of $\boldsymbol{\gamma}_0$ and $\boldsymbol{\phi}_0$ respectively. Then, We take $\hat{\boldsymbol{\beta}}$ as a root of $U(\boldsymbol{\beta},\ \hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\phi}}) = 0$, where

$$\begin{aligned} U(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\phi}) &= \sum_{i=1}^{N}\Big\{\frac{R_{iM}}{\pi_{iM}}\frac{\partial\boldsymbol{\mu}_i^T}{\partial\boldsymbol{\beta}}V_i(\boldsymbol{\rho})^{-1}(\boldsymbol{Y}_i-\boldsymbol{\mu}_i) \\ &\quad+\sum_{j=1}^{M-1}\frac{C_{ij}-\lambda_{i,j+1}R_{ij}}{\pi_{i,j+1}}\frac{\partial\boldsymbol{\mu}_i^T}{\partial\boldsymbol{\beta}}V_i(\boldsymbol{\rho})^{-1}(\boldsymbol{M}_{ij}-\boldsymbol{\mu}_i)\Big\}. \end{aligned}$$

By taylor expansion and law of large numbers, we have

$$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) = E\left(-\frac{1}{N}\frac{\partial U(\boldsymbol{\beta}_0,\ \hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\phi}})}{\partial\boldsymbol{\beta}^T}\right)^{-1}\frac{U(\boldsymbol{\beta}_0,\ \hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\phi}})}{\sqrt{N}}+o_p(1), \tag{2.5}$$

$$\begin{aligned} \frac{U(\boldsymbol{\beta},\ \hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\phi}})}{\sqrt{N}} &= \frac{U(\boldsymbol{\beta},\boldsymbol{\gamma}_0,\boldsymbol{\phi}_0)}{\sqrt{N}}+E\left(-\frac{\partial U(\boldsymbol{\beta},\ \boldsymbol{\gamma}_0,\ \boldsymbol{\phi}_0)}{\partial\boldsymbol{\gamma}}/N\right)\sqrt{N}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0) \\ &\quad+E\left(-\frac{\partial U(\boldsymbol{\beta},\ \boldsymbol{\gamma}_0,\ \boldsymbol{\phi}_0)}{\partial\boldsymbol{\phi}}/N\right)\sqrt{N}(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi}_0)+o_p(1), \end{aligned} \tag{2.6}$$

where

$$\begin{aligned} \frac{\partial U(\boldsymbol{\beta},\ \boldsymbol{\gamma},\ \boldsymbol{\phi})}{\partial\boldsymbol{\gamma}} &= \sum_{i=1}^{N}\sum_{j=1}^{M-1}\frac{C_{ij}-\lambda_{i,j+1}R_{ij}}{\pi_{i,j+1}}\frac{\partial\boldsymbol{\mu}_i^T}{\partial\boldsymbol{\beta}}V_i(\boldsymbol{\rho})^{-1}\frac{\partial\boldsymbol{M}_{ij}}{\partial\boldsymbol{\gamma}}, \\ \frac{\partial U(\boldsymbol{\beta},\ \boldsymbol{\gamma},\ \boldsymbol{\phi})}{\partial\boldsymbol{\phi}} &= \sum_{i=1}^{N}\sum_{j=1}^{M-1}\frac{\partial\boldsymbol{\mu}_i^T}{\partial\boldsymbol{\beta}}V_i(\boldsymbol{\rho})^{-1}(\boldsymbol{M}_{ij}-\boldsymbol{Y}_i) \\ &\quad\times\frac{\partial(C_{ij}-\lambda_{i,j+1}R_{ij})/\pi_{i,j+1}}{\partial\boldsymbol{\phi}}. \end{aligned}$$

(2.7) equals 0 if $\lambda_{ij}(\boldsymbol{\phi}_0) = P(R_{ij}=0\mid\bar{\boldsymbol{L}}_{i,j-1},\ R_{i,j-1}=1\ ;\ \boldsymbol{\phi}_0)$ for all i and j, and (2.8) equals 0 if $M_{ij}(\boldsymbol{\gamma}_0) = E\left[\boldsymbol{Y}_i\ \mid\ \bar{\boldsymbol{L}}_{ij},\ R_{ij}=1\ ;\ \boldsymbol{\gamma}_0\right]$ for all i and j.

Hence, if both the auxiliary models are correctly specified,

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = E\left(-\frac{1}{N}\frac{\partial U(\boldsymbol{\beta}_0,\ \hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\phi}})}{\partial \boldsymbol{\beta}^T}\right)^{-1}\frac{U(\boldsymbol{\beta}_0,\ \boldsymbol{\gamma}_0,\ \boldsymbol{\phi}_0)}{\sqrt{N}} + o_p(1). \qquad (2.7)$$

Therefore, the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is ABA^T, where A is inverse of probability limit of $N^{-1}\partial U(\boldsymbol{\beta}_0,\ \hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\phi}})/\partial\boldsymbol{\beta}^T$ and B is the aymptotic variance of $N^{-1/2}U(\boldsymbol{\beta}_0,\ \boldsymbol{\gamma}_0,\ \boldsymbol{\phi}_0)$, and it is consistently estimated by (2.4). However, if one of the auxiliary models is mis-specified, the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is not the same with ABA^T, and (2.4) is not consistent estimator of it anymore.

Chapter 3

Proposed Method

In this chapter, we propose alternative DR-GEE, which allow (2.4) to be consistent variance estimator of $\hat{\boldsymbol{\beta}}$. Let $(\hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\phi}})$ be a solution to equations

$$\left(\frac{\partial U(\boldsymbol{\beta},\ \boldsymbol{\gamma},\ \boldsymbol{\phi})}{\partial \boldsymbol{\gamma}},\ \frac{\partial U(\boldsymbol{\beta},\ \boldsymbol{\gamma},\ \boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right) \ = \ (0,\ 0). \tag{3.1}$$

Before we describe a theorem, we introduce some regularity conditions.

C 1 $\exists\ (\hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\phi}})$ *: unique solution to* $\partial U(\boldsymbol{\beta}_0,\ \boldsymbol{\gamma},\ \boldsymbol{\phi})/\partial(\boldsymbol{\gamma},\ \boldsymbol{\phi}) = 0$.

C 2 $\exists\ (\boldsymbol{\gamma}^\star,\ \boldsymbol{\phi}^\star)$ *is unique solution to* $E(\partial U(\boldsymbol{\beta}_0,\ \boldsymbol{\gamma},\ \boldsymbol{\phi})/\partial(\boldsymbol{\gamma},\ \boldsymbol{\phi})) = 0$ *and in the interior of its parameter space,* Ω.

C 3 $\partial U(\boldsymbol{\beta}_0,\ \boldsymbol{\gamma},\ \boldsymbol{\phi})/\partial(\boldsymbol{\gamma},\ \boldsymbol{\phi}) \ \xrightarrow{p} E(\partial U(\boldsymbol{\beta}_0,\ \boldsymbol{\gamma},\ \boldsymbol{\phi})/\partial(\boldsymbol{\gamma},\ \boldsymbol{\phi}))$ *uniformly in* Ω

C 4 $E(\partial U(\boldsymbol{\beta}_0,\ \boldsymbol{\gamma},\ \boldsymbol{\phi})/\partial(\boldsymbol{\gamma},\ \boldsymbol{\phi}))$ *is continuously differentiable on some neighborhood* $\Omega_0\ (\subset \Omega)$ *of* $(\boldsymbol{\gamma}^\star,\ \boldsymbol{\phi}^\star)$ *almost everywhere.*

C 5 $N^{1/2}\ \partial U(\boldsymbol{\beta}_0,\ \boldsymbol{\gamma},\ \boldsymbol{\phi})/\partial(\boldsymbol{\gamma},\ \boldsymbol{\phi}) \ \xrightarrow{d}\ N(0,\ A)$ *for some positive definite* A.

C 6 $\frac{\partial^2}{\partial(\boldsymbol{\gamma},\ \boldsymbol{\phi})\partial(\boldsymbol{\gamma},\ \boldsymbol{\phi})^T}U(\boldsymbol{\beta}_0,\ \boldsymbol{\gamma},\ \boldsymbol{\phi})$ *uniformly converges in probability to* $B(\boldsymbol{\gamma},\ \boldsymbol{\phi})$, *for some* $(\boldsymbol{\gamma},\ \boldsymbol{\phi})$ *continuous at* $(\boldsymbol{\gamma}^\star,\ \boldsymbol{\phi}^\star)$ *and* $B(\boldsymbol{\gamma}^\star,\ \boldsymbol{\phi}^\star)$ *is nonsingular.*

C 7 $N^{-1/2}\ \partial U(\boldsymbol{\beta}_0,\ \boldsymbol{\gamma},\ \boldsymbol{\phi})/\partial(\boldsymbol{\gamma},\ \boldsymbol{\phi}) = o_p(1)$.

Theorem 1 *Suppose (C 1) - (C 7) hold, then $\hat{\boldsymbol{\beta}}$ satisfies doubly robustness, and formula (2.4) is consistent variance estimator of $\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the unique solution to $U(\boldsymbol{\beta}_0,\ \hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\phi}})$.*

Proof

(C 1) - (C 7) implies that $(\hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\phi}}) = (\boldsymbol{\gamma}^\star,\ \boldsymbol{\phi}^\star) + o_p(1)$ and $N^{1/2}((\hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\phi}}) - (\boldsymbol{\gamma}^\star,\ \boldsymbol{\phi}^\star)) = O_p(1)$. If the droupout model is correctly specifid, then $\boldsymbol{\phi}_0$ is a solution to $E(\partial U(\boldsymbol{\beta}_0,\ \boldsymbol{\gamma}^\star,\ \boldsymbol{\phi})/\partial\boldsymbol{\gamma}) = 0$, so that $\boldsymbol{\phi}_0 = \boldsymbol{\phi}^\star$ and $\hat{\boldsymbol{\beta}}$ is consistent. If the imputation model is correctly specified, then $\boldsymbol{\gamma}_0$ is a solution to $E(\partial U(\boldsymbol{\beta}_0,\ \boldsymbol{\gamma},\ \boldsymbol{\phi}^\star)/\partial\boldsymbol{\phi}) = 0$, so that $\boldsymbol{\gamma}_0 = \boldsymbol{\gamma}^\star$ and $\hat{\boldsymbol{\beta}}$ is consistent. Hence, $\hat{\boldsymbol{\beta}}$ satisfies doubly robustness.

Moreover, $(\hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\phi}}) = (\boldsymbol{\gamma}^\star,\ \boldsymbol{\phi}^\star) + o_p(1)$ and $N^{1/2}((\hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\phi}}) - (\boldsymbol{\gamma}^\star,\ \boldsymbol{\phi}^\star)) = O_p(1)$ imply

$$E\left(-\frac{\partial U(\boldsymbol{\beta},\ \boldsymbol{\gamma}^\star,\ \boldsymbol{\phi}^\star)}{\partial\boldsymbol{\gamma}}/N\right)\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^\star) = o_p(1), \tag{3.2}$$

$$E\left(-\frac{\partial U(\boldsymbol{\beta},\ \boldsymbol{\gamma}^\star,\ \boldsymbol{\phi}^\star)}{\partial\boldsymbol{\phi}}/N\right)\sqrt{N}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^\star) = o_p(1). \tag{3.3}$$

Therefore, by (2.5), (2.6), (3.2) and (3.3)

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = E\left(-\frac{1}{N}\frac{\partial U(\boldsymbol{\beta}_0,\ \boldsymbol{\gamma}^\star,\ \boldsymbol{\phi}^\star)}{\partial\boldsymbol{\beta}^T}\right)^{-1}\frac{U(\boldsymbol{\beta}_0,\ \boldsymbol{\gamma}^\star,\ \boldsymbol{\phi}^\star)}{\sqrt{N}} + o_p(1),$$

i.e $\hat{\boldsymbol{\beta}}$ is asymptotically independent of $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}})$. Hence, formula (2.4) is consistent variance estimator of $\hat{\boldsymbol{\beta}}$.

Chapter 4

Simulation Study

We conducted simulation studies to compare bias and coverage probability of WGEE, original DR-GEE and alternative DR-GEE. performance of our proposed estimator with the two estimators introduced in chapter 2. We set the sample size N=3000, and the maximum number of time points m=3. As for the covariate, we set $\boldsymbol{X}_{ij} = (1,\ x_i)$, where x_i is treatment indicator which is 1 if it belongs to treatment group, and -1 otherwise. we use 3 types of true covariance matrix Σ.

$$\begin{pmatrix} 1 & 0.4 & 0.4 \\ 0.4 & 1 & 0.4 \\ 0.4 & 0.4 & 1 \end{pmatrix},\ \begin{pmatrix} 1 & 0.4 & 0.4^2 \\ 0.4 & 1 & 0.4 \\ 0.4^2 & 0.4 & 1 \end{pmatrix},\ \begin{pmatrix} 1 & 0.3 & 0.4 \\ 0.3 & 1 & 0.3 \\ 0.4 & 0.3 & 1 \end{pmatrix}.$$

As for working correlation matrix, we use independence working correlation matrix. True imputation models and specified imputation models are set to be

$$\begin{aligned} E\left(Y_{i2} \mid \bar{\boldsymbol{L}}_{i1},\ R_{i1} = 1\right) &= \gamma_{210} + \gamma_{211} x_i + \gamma_{212} y_{i1} := y_{i2}^*, \\ E\left(Y_{i3} \mid \bar{\boldsymbol{L}}_{i2},\ R_{i2} = 1\right) &= \gamma_{320} + \gamma_{321} x_i + \gamma_{322} y_{i1} + \gamma_{323} y_{i2}, \\ E\left(Y_{i3} \mid \bar{\boldsymbol{L}}_{i2},\ R_{i2} = 1\right) &= \gamma_{320} + \gamma_{321} x_i + \gamma_{322} y_{i1} + \gamma_{323} y_{i2}^*. \end{aligned}$$

True observation probability models are

$$\begin{aligned} logit\ P(R_{i2}=0|\bar{\boldsymbol{L}}_{i1},\ R_{i1}=1) &= \phi_{10}+\phi_{11}x_i+\phi_{12}y_{i1}+\phi_{13}x_iy_{i1}, \\ logit\ P(R_{i3}=0|\bar{\boldsymbol{L}}_{i2},\ R_{i2}=1) &= \phi_{20}+\phi_{21}x_i+\phi_{22}y_{i1}+\phi_{23}y_{i2} \\ &\quad +\phi_{24}x_iy_{i1}+\phi_{25}x_iy_{i2}+\phi_{26}y_{i1}y_{i2}, \end{aligned}$$

While Specified observation probability model

$$\begin{aligned} logit\ P(R_{i2}=0|\bar{\boldsymbol{L}}_{i1},\ R_{i1}=1) &= \phi_{10}+\phi_{11}x_i+\phi_{12}y_{i1}, \\ logit\ P(R_{i3}=0|\bar{\boldsymbol{L}}_{i2},\ R_{i2}=1) &= \phi_{20}+\phi_{21}x_i+\phi_{22}y_{i1}+\phi_{23}y_{i2}. \end{aligned}$$

It means one auxiliary model is misspecifid. True $\boldsymbol{\beta}$ is $(1,\ 1)^T$, true $\boldsymbol{\phi}_1$ is $(0.3,\ 0.2,\ 0.2,\ 15.0)^T$ and true $\boldsymbol{\phi}_2$ is $(0.3,\ 0.2,\ 0.2,\ 0.2,\ 15.0,\ 15.0,\ 5.0)^T$. The simulation results are presented in table (4.1). The proposed DR-GEE achieved coverage probability close to 0.95, while the others did not. For these results, the variance estimator of DR-GEE proposed is consistent, while the others are not.

Table 4.1: Simulation estimates of bias, and coverage probability(C.P.) when the model for probability of missingness is misspecified.

working		$\hat{\beta}_0$		$\hat{\beta}_1$	
correlation	method	bias	C.P.	bias	C.P.
intraclass	WGEE	0.0003	0.941	-0.0149	0.924
	DR GEE	0.0004	0.918	0.0007	0.925
	DR GEE proposed	0.0004	0.941	0.0008	0.945
AR1	WGEE	0.0032	0.956	-0.0122	0.918
	DR GEE	0.0033	0.933	-0.0004	0.906
	DR GEE proposed	0.0033	0.957	-0.0004	0.937
unstructured	WGEE	-0.0008	0.948	-0.0144	0.912
	DR GEE	-0.0008	0.943	-0.0006	0.918
	DR GEE proposed	-0.0008	0.949	-0.0007	0.936

Chapter 5

Application to data from clinical trial in antihypertensive drug

The data we analyzed is from clinical trial in antihypertensive drug. It is clinical trial of treatment for isolated systolic hypertension. The trial is randomized, double blind, and placebo-controlled trial. The number of study participants is 4736, of which 63% are men, and 37% are women. Age of participants is 60 years or older. They were randomized to active treatment or placebo. We are interested in the treatment effect on mean arterial pressure (weighted average of systolic blood pressure and diastolic blood pressure). Responses were repeatedly measured for 3 times. In this paper, the full data were not used. We only used partial data. Covariates are group indicator, age and time point. We computed $\hat{\boldsymbol{\beta}}$ and its standard error.

Table 5.1: Comparing point estimates and its standard error (S.E.) of 3 methods.

Method	RZGRP		AGE		time	
	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.
WGEE	5.18	1.583	-0.24	0.148	-0.39	0.532
DR GEE	5.43	1.438	-0.24	0.125	-0.39	0.476
DR GEE proposed	5.32	1.409	-0.23	0.122	-0.39	0.469

Chapter 6

Discussion

The proposed DR-GEE method has a computational issue. When the number of covariates in auxiliary models is large, the number of estimating equations for auxiliary models is much larger than the number of parameters of it. In this case, generalized method of moment is applicable. The method is extremely slow in computation and may not coverge. Alternative algorithm for this problem need to be constructed.

Other problem is small sample bias. DR-GEE is based on asymptotic theories and so be proposed one. When the models are complicated, the proposed method need huge sample size for its asymptotic properties to work. Another method to reduce small sample bias need to be researched.

References

Liang, K.Y., and Zeger, S.L. (1986), "Longitudinal data analysis using generalised linear models," *Biometrics*, **73**, 13–22.

Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1995), " Analysis of semiparametric regression models for repeated outcomes in the presence of missing data.," *Journal of the American Statistical Association*, **90**, 106–121.

Paik, M.C. (1997), "The generalized estimating equations approach when data are not missing completely at random," *Journal of the American Statistical Association*, **92**, 1320–1329.

Seaman, S. and Copas, A. (2009), "Doubly robust generalized estimating equations for longitudinal data," *Statistics in Medicine*, **28**(2), 937–955.

Kim, J.K., and Shao, J. (2014), " Statistical Methods for Handling Incomplete Data ," *CRC press.* 6–7, 112–114.

Tsiatis, A.A. (2006), "Semiparametric Theory and Missing Data," *Springer*, 21–51.

국문초록

이중 로버스트성 일반화 추정 방정식 (DR-GEE) 은 반복 측정 자료에 결측치가 있는 경우 널리 사용되는 분석방법이다. 이 방법을 쓰기 위해서는 결과치와 관측 지시 변수에 대한 보조 모형을 가정해야 하며, 둘 중 하나의 모형이 정설정된 경우에는 일치 점 추정량을 얻을 수 있다. 이 방법의 또다른 장점은 간단한 분산 추정식인데, 이는 관심 모수의 추정치와 보조 모형의 추정치가 점근적으로 독립인 사실에 기인한다. 그러나, 이것은 두 보조 모형이 정설정된 경우에만 성립한다. 본 논문에서는 한 보조 모형이 오설정된 경우에도 분산의 일치 추정량을 얻을 수 있는 다른 방식의 DR-GEE를 제안한다. 주요 구상점은 관심 모수의 추정량과 보조 모형의 추정량이 점근적으로 독립이 되도록 보조 모형의 추정 방정식을 구성하는 것이다. 마지막으로, 이 방법으로 고혈압에 관한 임상연구 자료를 분석하는 예를 설명한다.

주요어 : 이중 로버스트성, 일반화 추정 방정식, 임의 결측, 분산의 일치 추정량.

학 번 : 2014-20300