Topics on Estimating Equations Approaches for Longitudinal Binary Outcomes with Report Bias

by

Chao Li

(Under the Direction of Ye Shen and Kevin K. Dobbin)

Abstract

Cocaine use is an important public health problem in the United States and throughout the world. It is associated with many medical consequences and psychosocial characteristics. Cognitive behavioral therapy (CBT) is an effective counseling intervention for supporting cocaine-dependent individuals through recovery and relapse prevention, or reducing their cocaine use by improving patient's motivation and enabling them to recognize risky situations. Our motivating example from the Self-reported Cocaine use with Urine test (SCU) data was based on a study of the effect of Cognitive behavioral therapy (CBT) on cocaine dependence at the Primary Care Center of Yale-New Haven Hospital. To evaluate the impact of adding CBT to physician management on cocaine dependent patients receiving buprenorphone, patients were randomly assigned to the treatment group and the control group. Collected outcomes included self-reported daily drug uses and weekly urine test results.

To date, Generalized Estimating Equations (GEE) are considered to be a reasonable approach to analyze the data with repeated measures binary outcomes. However, due to the existence of report bias in self-reported daily drug use, a direct application of GEE may not be valid for the SCU data. On the other hand, the less frequently measured urine test is considered more accurate. Therefore, we proposed Mean Corrected Generalized Estimating Equations (MCGEE) to estimate the treatment effect in self-reported binary outcomes. The urine test is used to detect the contamination and correct the model's mean in the equation. We demonstrated that the proposed approach yield consistent and asymptotically normally distributed estimators with unbiased contamination probability. However, we also noticed that when the time period for cocaine to be cleared from urine increased, bias of the estimators of the MCGEE approach increased. Thus, we proposed to include a weight function of the contamination probability into the MCGEE and build Mean Corrected Weighted Generalized Estimating Equations (MCWGEE) to further control the potential bias of the estimators. Additionally, we also investigated the impacts of patients' dropouts in the SCU data using MCWGEE with an extra weight from the estimated probability of dropout at the time of attrition.

INDEX WORDS: GEE, WGEE, Report bias, Cocaine use, Longitudinal Binary Data

Topics on Estimating Equations Approaches for Longitudinal Binary Outcomes with Report Bias

by

Chao Li

M.P.H., University of Georgia, 2012

M.S., University of Georgia, 2017

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2017

Topics on Estimating Equations Approaches for Longitudinal Binary Outcomes with Report Bias

by

Chao Li

Major Professors:	Ye Shen Kevin K. Dobbin
Committee:	Stephen L. Rathbun Zhenqiu(Laura) Lu

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2017

Acknowledgments

I would like to thank my advisor Dr. Shen for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me throughout my research and the writing of this dissertation. I could not have imagined having a better advisor and mentor for my Ph.D study. Without his counsel and encouragement, I could not finish such a complex work.

Also, I want to thank my co-advisor Dr. Dobbin, and committee members, Dr. Rathbun, and Dr. Lu for their help on reviewing my work and giving valuable suggestions. Moreover, I would like to appreciate the faculty members and staff from the Department of Epidemiology and Biostatistics; they provided constant help during my years at The University of Georgia.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Report bias is an important problem in survey research. It occurs when a respondent's answer in the survey differs from the true value for that respondent (Del Boca, Noll 2000). Although there are various benefits of self-reported data, such as: efficiency, convenience, adaptability, flexibility, and relatively low cost (Del Boca and Noll 2000); many factors may influence the quality of self-reported data and cause bias. Sources of report bias related to drug use include participant has difficulty understanding the questions; respondent has trouble recalling the information needed to answer the questions; participant is not willing to report, social pressures, etc (Johnson and Fendrich 2005). Report bias may reduce the precision of the estimation of dependent variables and violate statistical inferences (Biemer and Trewin, 1997).

Our motivating example from the Self-reported Cocaine use with Urine test (SCU)

data was based on a study of the effect of Cognitive behavioral therapy (CBT) on cocaine dependence at the Primary Care Center of Yale-New Haven Hospital. To evaluate the impact of adding CBT to physician management on cocaine dependent patients receiving buprenorphone, patients were randomly assigned to the treatment or the control group. Collected outcomes included self-reported daily drug use and weekly urine test results. To date, Generalized Estimating Equations (GEE) are considered to be a reasonable approach to analyze data with repeated measured outcome. However, due to the existence of report bias in self-reported daily drug use, a direct application of GEE may not be valid in the SCU data. On the other hand, the less frequently measured urine test result was considered more accurate. Therefore, we proposed adjusting the GEE model through the corrected marginal means based on the detected report bias in the self-reported daily data using the urine test results.

The dissertation is organized as follows. In Chapter 2, we propose Mean Corrected Generalized Estimating Equations (MCGEE) to estimate the treatment effect in self-reported data. The urine test was used to detect the contamination and correct the model's mean in the equation. However, we also noticed that when the time period for cocaine to be cleared from urine increased, the bias of MCGEE estimators increased. Thus, we proposed including a weight function of the contamination probability to construct Mean Corrected Weighted Generalized Estimating Equations (MCWGEE) to further reduce the potential bias of the estimators, and described the approach in Chapter 3. Additionally, we also investigated the impacts of patients' dropouts in the SCU data using MCWGEE with an extra weight from the estimated probability

of dropout at the time of attrition in Chapter 4.

Chapter 2

Generalized Estimating Equations Approach for Longitudinal Binary Outcomes with Report Bias

2.1 Introduction

Cocaine use is an important public health problem in the United States and throughout the world. It is associated with many medical consequences and psychosocial characteristics, including increased risk of myocardial infarction, stroke, infectious diseases, chronic stress and violence (Macdonald et al., 2008, Qureshi et al., 2001). Despite its negative impact on health, there is no effective pharmacological therapy specifically targeted for cocaine addiction (Sofuoglu and Kosten, 2006).

Although no medications are currently available to treat cocaine addiction effectively, one promising substitute for cocaine is buprenorphine, a partial mu-opioid agonist at the mu-opioid receptor and kappa-opioid antagonist. Its efficacy has been noted in some pharmacotherapy trials (Brown et al. 1991; Kamien et al. 1991). A recent study showed that the combination of buprenorphine, naloxone, and naltrexone may reduce cocaine use among subjects addicted to cocaine as well as past or current opioid dependents (Ling et al. 2016). The cocaine use reduction with buprenorphine randomized clinical trial, conducted by the National Institute on Drug Abuse Clinical Trials Network, demonstrated that buprenorphine reduced cocaine use in adults with cocaine dependence and opioid use disorders (Mooney et al. 2013).

Cognitive behavioral therapy (CBT) is an effective counseling intervention for drug and alcohol use disorders which includes learning skills and strategies for regulating effect, changing maladaptive thoughts, and learning new behavioral strategies (Magill and Ray 2009). It has proven to be effective in supporting cocaine-dependent individuals through recovery and relapse prevention, or reducing their cocaine use by improving patient's motivation and enabling them to recognize risky situations (Maude-Griffin et al., 1998, Gonzalez et al., 2006). Following therapy, the patient learns to identify thoughts, feelings, and events that precede and follow each time of cocaine use and to develop and rehearse coping skills (Beck et al., 1993). Maude-Griffin et al. (1998) evaluated the efficacy of CBT among cocaine dependents in a randomized clinical trial, and found a significant number of participants in CBT had abstained from cocaine use. A meta-analysis examined 53 randomized controlled

trials of CBT for adults diagnosed with alcohol or drug use disorders; showed that CBT had produced a statistically significant treatment effect (Magill and Ray 2009). Rawson et al. (2006) conducted a randomized clinical trial to compare contingency management and CBT for stimulant-dependent individuals, and concluded that CBT reduced drug use from baseline levels to all measures at follow-ups.

Our motivating data, Self-reported Cocaine use with Urine test (SCU), is based on a study of the effect of CBT on cocaine dependence at the Primary Care Center of Yale-New Haven Hospital. To evaluate the impact of adding CBT to physician management in cocaine dependent patients receiving buprenorphone, patients were randomly assigned to the treatment group and control group, both receiving buprenorphone that is stored in bottles. The control group receives physical management (PM), a 15-20 minutes session by Internal Medicine physicians with experience administering buprenorphone. The treatment group receives PM and CBT, which is provided by trained clinicians. Collected outcomes include self-reported daily drug use and weekly urine test results. A recent publication analyzed the data set using a logistic regression model with the daily self-reported drug use as repeated outcomes, and found no significant effect from the CBT intervention (Fiellin et al. 2013). However, their analysis was based on the self-reported outcomes without considering the possibility of report bias.

Report bias, sometimes referred as measurement error, occurs when a respondent's answer in the survey differs from the true value of that respondent (Del Boca, Noll

2000). It represents a major problem in the assessment of self-reported data, which is commonly collected in clinical research of intervention on drug use. The benefits of self-reported data are: efficiency, convenience, adaptability, flexibility, and relatively low cost (Del Boca and Noll 2000). However, many factors may influence self-reported data and cause bias. The sources of report bias related to drug use may include: participant has difficulties in understanding the questions; respondent has issues in recalling the information needed to answer the questions; participant is not willing to report, social pressures, etc (Johnson and Fendrich 2005). Report bias may reduce the precision of the parameters estimation and violate causal inferences (Biemer and Trewin, 1997).

In studies of participants with drug use dependency, self-reported drug use outcomes are subject to report bias. Urine test result is sometimes used as a surrogate marker, as it is usually more accurate and reliable. However, the collection of a biological test is more expensive and less convenient compared to self-report (Magura et al., 1987). In practice, urine test is often collected less frequently than self-reported data. With the limitations of self-reported data being its inaccuracy and of urine indicators being its infrequency, it has been suggested to either combine these two measures as joint outcomes, or use biological test results to correct the bias in self-reports to increase their validity (Babor et al., 2000, Blattman et al. 2016, Wilcox et al. 2013).

Wilcox et al. (2013) examined the concordance between self-reported drug use data and urine test results among adolescents and young adults with opioid dependence

participating in a clinical trial. They used a generalized linear mixed model, and concluded that the concordance between these two results was reasonably high. The Kappa coefficient is often used as a measurement of the agreement between self-reported data and biological test results. Babor et al. repeatedly collected self-reported drinking outcomes and measured them on their biological indicators. Based on the Kappa coefficient, the correspondence of these two measurements was 97.1% at the baseline of the clinical trial and decreased to 84.7% at the 15-month follow-up (Babor et al. 1997). Sherman and Bigelow (1992) reported similar findings, showing that the agreement was high between the interviews and same day urine test results, but the validity of self-reports dropped at the 4-week follow-up. Another study compared urine test results and self-reported drug use from subjects of methadone maintenance in Australia, and the calculated Kappa coefficient didn't show any significant differences between these two outcomes (Digiusto et al. 1996). Winhusen et al. (2003) investigated the outcomes between self-reports and urine toxicology in cocaine clinical trials, and reported the correlation was around 0.40. While the reported concordance between the two outcomes is generally high, there are still significant portions of self-reports that have non-negligible discrepancies, especially at follow-ups in longitudinal studies.

When self-reports differ from urine test results, the cause of the discrepancy is usually due to a misunderstanding of the question or socially undesirability to answer. To assess factors that may affect self-reported measurements among alcohol and drug abuse patients by comparing blood and urine tests, there was a 97% agreement be-

tween verbal report and laboratory data for alcohol and 93% between verbal report and laboratory data for cocaine (Brown et al. 1992). Another study characterizing patients of cocaine dependence as under-reporters and truthful reporters based on their self-reports and urine test results, revealed that under-reporters attended more study sessions and were more likely to complete the study, and there were also significant differences in cocaine use patterns (Myrick et al. 2002). Babor et al. found that alcoholics who showed discrepancy in self-reports and blood test tended to drink more severely, and had more previous treatments and higher levels of cognitive impairment (Babor et al. 1997). A review on the validity of self-reports of alcohol consumption concluded that social context factors, respondent characteristics, and task attributes may influence self-reports (Del Boca and Darkes, 2003).

Most drug addiction studies with repeatedly self-reported outcomes used Generalized Estimating Equations (GEE) to analyze the data. Since its establishment in 1980s by Liang and Zeger (1986), GEE has been widely used for analysis of longitudinal data with repeated measurements. GEE is used to extend Generalized Linear Models (GLM) to a hierarchical setting with dependent outcomes by specifying a working correlation matrix (Fitzmaurice et al. 2009). Parameter estimators from GEE are consistent if the marginal means of the outcomes are correctly specified, and such consistency is retained even when the covariance structure is misspecified (Liang and Zeger 1986).

There are a variety of common structures that may be appropriate to model the work-

ing correlation matrix: independent, exchangeable, autoregressive, and unstructured. However, GEE remains asymptotically unbiased under misspecification of the correlation matrix, and provides consistent estimates (Liang and Zeger 1986). On the other hand, as GEE is an estimating procedure that relies on quasi-likelihood theory; the common properties of the likelihood approaches do not apply. As a result, the usual goodness-of-fit statistics cannot be easily derived (Wedderburn 1974).

In clinical studies, we are often confronted with projects that require the collection of longitudinal data with repeatedly measured binary outcomes. There are natural correlations among observations from the same subjects. GEE has been widely used to model such data. Umbricht et al. (2014) analyzed methadone effect on cocaine dependents in a double-blind randomized trial. The repeated measures of binary outcomes were analyzed using GEE with an autoregressive correlation structure, and they concluded that there was no significant difference in cocaine abstinence between the treatment group and the control group. GEE has also been used to examine the socio-demographic and behavioral factors associated with illicit substance injection in an observational cohort study; it suggested that several factors were significantly related to drugs injection (Lioyd-Smith et al. 2010). Another study performed GEE in a randomized clinical trial evaluating the effect of selective norepinephrine reuptake inhibitor on the cocaine dependence. The GEE analysis of the patients' urine samples revealed no significant differences between the treatment and the control groups (Walsh et al. 2012).

In the SCU Data, the direct application of the GEE may not be valid due to the potential report bias in self-reported daily drug use. Throughout the dissertation, such report bias will also be referred as "contamination". The urine test result is considered more accurate but is measured weekly, and the time period for cocaine to be cleared from urine can be longer than one day. In our preliminary analysis, urine results were used to detect report bias in the self-reported daily data. The results showed that there was roughly 21.4% contamination in self-reported outcomes. Since GEE can provide a consistent estimate if the marginal means of the outcomes are correctly specified, we used the urine test result to detect the contamination, and estimated the true marginal means of the self-reported results. This estimation was used in GEE to correct the report bias in the self-reported data in estimating the treatment effect.

The purpose of this chapter is to develop Mean Corrected Generalized Estimating Equations (MCGEE) for longitudinal datasets with report bias in binary outcome. The chapter is organized as follows. In section 2.2, we gave the notation and model equations. Section 2.3 studied the asymptotic properties of the estimators from our proposed approach. We further explored the bias of the estimators when data are contaminated in section 2.4. The performance of MCGEE on finite sample data were accessed through simulation studies in section 2.5. In section 2.6, we analyzed the Self-reported Cocaine use and Urine test (SCU) data using our proposed methods. Finally, the chapter is concluded with a discussion in section 2.7.

2.2 Methods

Key variables collected from the SCU dataset included daily self-reported cocaine use, weekly urine test results, and an indicator of treatment or control group status. The purpose of our study is to assess the treatment effect of CBT on the self-reported drug use outcomes.

2.2.1 True drug use

Let Y_{it} denote the true drug use variable, and X_{it} be the covariate vectors for estimation at times $t = 1, ..., T$ for subjects $i = 1, ..., N$. For the ith subject at time t, $Y_{it} = 1$ if the subject uses the drug, $Y_{it} = 0$ otherwise. The outcome Y_{it} is a binary response variable and its joint distribution is Bernoulli:

$$f_y\left(y_i \mid X_i\right) = pr\left(Y_{i1} = y_1, ..., Y_{iT} = y_T \mid X_i\right) = exp(y_{it}\eta_{it} - log(1 + exp(\eta_{it}))).$$

The marginal mean of the true drug use for the ith subject at a given time point t is denoted by μ_{it}. Let β be a vector of the regression parameters, then

$$\mu_{it} = E\left(Y_{it} \mid X_i, \beta\right) = Pr\left(Y_{it} = 1 \mid X_i, \beta\right),$$

and logit link function will be used

$$\eta_{it} = \log \frac{\mu_{it}}{1 - \mu_{it}} = x_{it}\beta.$$

Liang and Zeger(1986) have proposed GEE in the form

$$U_\beta(\beta) = \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it}' V_{it}^{-1} (Y_{it} - \mu_{it}) = 0,$$

where $D_{it} = \partial \mu_{it} / \partial \beta$ and V_i is the covariance matrix of Y_i, which can be decomposed into the form $A_i^{\frac{1}{2}} C_i(\gamma) A_i^{\frac{1}{2}}$, where A_i is a matrix with the marginal variances on the main diagonal and zeros elsewhere, γ is a vector which fully characterizes $C_i(\gamma)$, which serves as a working correlation matrix of the Y_i's.

After collecting μ_{it} in a vector $\mu_i = (\mu_{i1}, ...\mu_{iT})'$, and since we assumed $\eta_i = \log \frac{\mu_i}{1-\mu_i} = x_i\beta$,

$$D_i = \partial \mu_i / \partial \beta = \frac{e^{X_i'\beta}}{(1 + e^{X_i'\beta})^2} X_i.$$

With $A_i = diag(var(Y_{i1}), ..., var(Y_{iT}))$, $var(Y_{it}) = \mu_{it} \times (1 - \mu_{it}) = \frac{e^{x_{it}\beta}}{(1+e^{x_{it}\beta})^2}$, and

$$D_i = \partial \mu_i / \partial \beta = A_i X_i.$$

Hence, we can write the GEE in the form:

$$U_\beta(\beta) = \sum_{i=1}^{N} X_i' A_i (A_i^{\frac{1}{2}} C_i(\gamma) A_i^{\frac{1}{2}})^{-1} (Y_i - \mu_i) = 0.$$

Some common correlation structures for longitudinal data include: Exchangeable: correlation of two different time points of a subject $Y_{ij}, Y_{it} (j \neq t)$ is γ, Autoregressive:

correlation of $Y_{ij}, Y_{it}(j \neq t)$ is $\gamma^{|j-t|}$, and Unstructured: correlation of $Y_{ij}, Y_{it}(j \neq t)$ is γ_{jt}.

If we assume the correlation matrix, $C_i(\gamma)$, of Y_i to be an Identity matrix, the estimating equation can be simplified as:

$$\begin{aligned} U_\beta(\beta) &= \sum_{i=1}^{N} X_i' A_i (A_i^{\frac{1}{2}} I A_i^{\frac{1}{2}})^{-1} (Y_i - \mu_i) \\ &= \sum_{i=1}^{N} X_i' (Y_i - \mu_i) = 0. \end{aligned}$$

The GEE equation is reduced to the score equation from the likelihood approach. The solution of this equation is the same as the maximum likelihood estimate (MLE) of β.

Assuming that the marginal mean μ_i has been correctly modeled, the estimator $\hat{\beta}$ is normally distributed with the mean being equal to β and the covariance matrix:

$$Var\left(\hat{\beta}\right) = I_0^{-1} I_1 I_0^{-1},$$

where

$$I_0 = \left(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{D}_i\right),$$

and

$$I_1 = \left(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{A}_i \hat{V}_i^{-1} \hat{D}_i\right).$$

The solution $\hat{\beta}$ can be obtained by a Fisher's scoring algorithm, which first gives an initial guess for $\hat{\beta}^0$, and then updates $\hat{\beta}^l$ in the lth iteration by taking:

$$\hat{\beta}^{l+1} = \hat{\beta}^l - \left(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{D}_i \right)^{-1} \left(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \left(Y_i - \hat{\mu}_i \right) \right).$$

The form of the robust variance estimator (sandwich estimator) for $\hat{\beta}$ is:

$$\left(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{D}_i \right)^{-1} \left(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{A}_i \hat{V}_i^{-1} \hat{D}_i \right) \left(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{D}_i \right)^{-1}.$$

This estimate is consistent even if V_i, the covariance matrix of Y_i is misspecified.

The true drug use variable Y_{it} cannot be observed in our study. In practice, self-reported data is directly used to replace Y_{it}. The discrepancy between the two may cause a significantly biased estimation of β if we directly apply GEE. Since the GEE only requires its marginal mean μ_i to be correctly specified to provide a consistent estimate of β without the need of a correct working correlation $C_i(\gamma)$, we propose Mean Corrected Generalized Estimated Equations (MCGEE) by adjusting the marginal mean of self reported outcomes based on the urine test result to estimate the treatment effect.

2.2.2 Self-reported drug use

Let R_{it} represent an indicator variable for outcome contamination at times $t = 1, ..., T$, for subjects $i = 1, ..., N$, suggesting whether self-reported data is the same as true drug use. Set $R_{it} = 1$ if there exists contamination, i.e. self reported data is not the same as true drug use data, otherwise $R_{it} = 0$. Let Z_{it} denote the self reported drug use, then

$$Z_{it} = Y_{it}\left(1 - R_{it}\right) + \left(1 - Y_{it}\right) R_{it}.$$

To estimate μ^*_{it}, the expected value of Z_{it}, we assume that the true drug use variable Y_{it} and the indicate variable for contamination R_{it} are independent given the covariate X_{it}. Then, μ^*_{it} can be calculated as:

$$\begin{aligned}\mu^*_{it} &= E(Z_{it}|X_{it}, \beta) \\ &= E(Y_{it}|X_{it}, \beta) \times E((1 - R_{it})\,|X_{it}, \beta) + E((1 - Y_{it})\,|X_{it}, \beta) \times E(R_{it}|X_{it}, \beta).\end{aligned}$$

Since $E(Y_{it}|X_{it}, \beta) = \mu_{it}$,

$$\begin{aligned}\mu^*_{it} = E(Z_{it}|X_{it}, \beta) &= \mu_{it} \times E((1 - R_{it})\,|X_{it}, \beta) + (1 - \mu_{it}) \times E(R_{it}|X_{it}, \beta) \\ &= \mu_{it} - 2\mu_{it} \times E(R_{it}|X_{it}, \beta) + E(R_{it}|X_{it}, \beta).\end{aligned}$$

We assume that $E(R_{it}|X_{it},\beta) = p_{it}$, and $p_i = (p_{i1}, ...p_{iT})'$. After collecting μ^*_{it} in a vector $\mu^*_i = (\mu^*_{i1}, ...\mu^*_{iT})'$,

$$\mu^*_i = E(Z_i|X_i,\beta),$$

where its tth element is:

$$\mu^*_{it} = \mu_{it} - 2\mu_{it} \times p_{it} + p_{it}.$$

The MCGEE form of the self-reported data Z_i is defined as:

$$U^*_\beta(\beta) = \sum_{i=1}^{N} D^{*\prime}_i V^{*-1}_i (Z_i - \mu^*_i) = 0,$$

$$\begin{aligned} D^*_i = \partial\mu^*_i/\partial\beta &= (1-2p_i) \otimes \frac{\partial\mu_i}{\partial\beta} \\ &= (1-2p_i) \otimes A_i X_i, \end{aligned}$$

$A_i = diag(var(Y_{i1}), ..., var(Y_{iT}))$, $var(Y_{it}) = \mu_{it} \times (1-\mu_{it}) = \frac{e^{x_{it}\beta}}{(1+e^{x_{it}\beta})^2}$, and $\otimes$ means only multiplying the rows of a vector to their corresponding rows of a matrix. In our framework, we multiply the tth row of vector $1-2p_i$ to the same tth row of matrix A_iX_i, i.e., $(1-2p_i) \otimes A_iX_i = ((1-2p_{i1}) \times a_{i1}x_{i1}, ..., (1-2p_{iT}) \times a_{iT}x_{iT})'$.

The matrix V^*_i is the covariance matrix of Z_i, which can be decomposed into the form $A^{*\frac{1}{2}}_i C^*_i(\gamma) A^{*\frac{1}{2}}_i$, where A^*_i is a matrix with the marginal variances on the main

diagonal and zeros elsewhere, i.e., $A_i^* = diag(var(Z_{i1}), ..., var(Z_{iT}))$, and

$$\begin{aligned}
var(Z_{it}) &= \mu_{it}^*(1-\mu_{it}^*) \\
&= (\mu_{it} - 2\mu_{it}p_{it} + p_{it})(1 - \mu_{it} + 2\mu_{it}p_{it} - p_{it}) \\
&= \mu_{it} - 2\mu_{it}p_{it} + p_{it} - \mu_{it}^2 + 2\mu_{it}^2 p_{it} - p_{it}\mu_{it} + 2\mu_{it}^2 p_{it} - 4\mu_{it}^2 p_{it}^2 + 2\mu_{it}p_{it}^2 - \mu_{it}p_{it} \\
&\quad + 2\mu_{it}p_{it}^2 - p_{it}^2 \\
&= (1-2p_{it})^2\mu_{it}(1-\mu_{it}) + p_{it}(1-p_{it}) \\
&= (1-2p_{it})^2 var(Y_{it}) + p_{it}(1-p_{it}).
\end{aligned}$$

The parameter γ is a vector which fully characterizes $C_i^*(\gamma)$, the working correlation matrix of the Z_i's.

Therefore, we can write the self-reported data's MCGEE in the form:

$$U_\beta^*(\beta) = \sum_{i=1}^{N}(1-2p_i) \otimes X_i' A_i (A_i^{*\frac{1}{2}} C_i^*(\gamma) A_i^{*\frac{1}{2}})^{-1} (Z_i - (\mu_i - 2\mu_i \otimes p_i + p_i)) = 0.$$

When the correlation matrix, $C_i^*(\gamma)$, of Z_i takes the form of an Identity matrix, the estimating equations are:

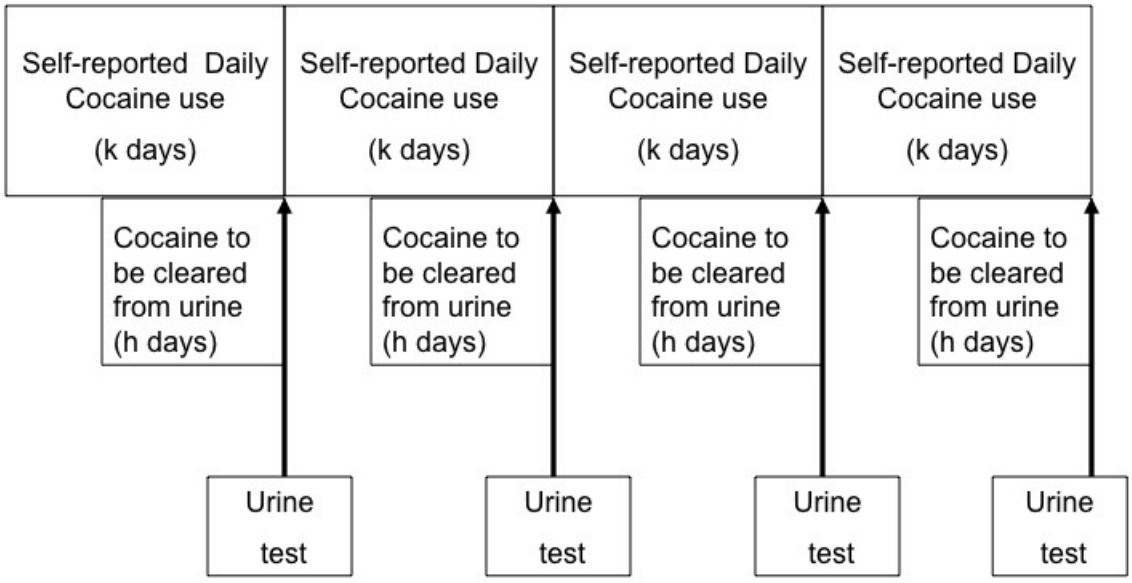
$$\begin{aligned}
U_\beta^*(\beta) &= \sum_{i=1}^{N}(1-2p_i) \otimes X_i' A_i A_i^{*-1} (Z_i - (\mu_i - 2\mu_i \otimes p_i + p_i)) \\
&= \sum_{i=1}^{N}(1-2p_i) \otimes X_i' A_i ((1-2p_i)^2 \otimes A_i + p_i \otimes (1-p_i))^{-1} (Z_i - (\mu_i - 2\mu_i \otimes p_i + p_i)) \\
&= 0.
\end{aligned}$$

Two approaches are considered to estimate the expected value of R_{it}. First, assume each subject has one single $E(R_{it}|X_{it}, \beta)$ at all time points, then the indicator variable R_{it} follows a Bernoulli Distribution with contamination probability p_i. Therefore, $E(R_{i1}) = ... = E(R_{it}) = p_i$, where $i = 1, ..., N$. And $(p_1, ..., p_N)^T$ is a vector of contamination probability for subjects 1 to N.

In our working dataset, urine samples are collected every k days, and the time period for cocaine to be cleared from urine is h days ($h \leq k$) (Figure 2.1). For the ith patient, U_{ij} denotes urine test results from subject i at the jth measurement, where $j = 1, ..., m_i$, m_i is the number of urine measurements for subject i. We divide the whole time period into multiple $k-day$ blocks, and calculate the sum of self-reported cocaine use over h days, $\sum_{t=k\times j-h+1}^{k\times j} Z_{it}$, for $j = 1, ..., m_i$. An indicator variable I_{ij} is defined as: $I_{ij} = 0$ if $\sum_{t=k\times j-h+1}^{k\times j} Z_{it} = 0$, otherwise $I_{ij} = 1$.

The difference between each urine test result U_{ij} and the indicator variable I_{ij} is calculated as: $F_{ij} = |U_{ij} - I_{ij}|$, where $i = 1, ..., N; j = 1, ..., m_i$. We then use this difference variable F_{ij} to detect contamination, in which case $F_{ij} = 0$ indicates that we failed to detect any contamination, while $F_{ij} = 1$ suggests a contamination detection in the block.

Figure 2.1: Data structure



Naturally, the contamination probability p_i can be estimated as:

$$\hat{p}_i = \frac{\sum_{j=1}^{m} F_{ij}}{m_i}.$$

However, this estimate assumes that the contamination probability for the first $k-h$ days in each time block is the same as the last h days, which may not be true. Moreover, even if we successfully detect contamination in a block, it is still challenging to determine the exact number and location of contaminations within the block. In some scenarios, the contamination probability may be underestimated.

On the other hand, to estimate the contamination probability for each observation, we assume that the contamination indicator R_{it} depends on some covariates and can be modeled through logistic regression models.

As defined earlier, $R_{it} = 1$ if self-reported drug use is not the same as true drug use, otherwise $R_{it} = 0$. Assume

$$\log \frac{Pr\left(R_{it} = 1 | X_i, B_{it}, \theta\right)}{1 - Pr\left(R_{it} = 1 | X_i, B_{it}, \theta\right)} = \theta_0 + \theta_1 X_i + \theta_2 B_{it},$$

so

$$p_{it} = E\left(R_{it} \mid X_i, B_{it}, \theta\right) = Pr\left(R_{it} = 1 | X_i, B_{it}, \theta\right) = \frac{e^{\theta_0 + \theta_1 X_i + \theta_2 B_{it}}}{1 + e^{\theta_0 + \theta_1 X_i + \theta_2 B_{it}}},$$

where X_i denotes a vector of time independent covariates, B_{it} denotes a vector of time dependent covariates, and θ represents a vector of the regression parameters.

To estimate the contamination probability using urine data, we fit a model using the difference between the urine test result and the indicator variable F_{ij}, which is calculated in the previous approach, time independent covariates X_i, and a function of time dependent covariates for h days B'_{ij}.

$$\log \frac{Pr\left(F_{ij}=1|X_i, B'_{ij}, \theta'\right)}{1-Pr\left(F_{ij}=1|X_i, B'_{ij}, \theta'\right)} = \theta'_0 + \theta'_1 X_i + \theta'_2 B'_{ij},$$

where $i = 1, ..., N; j = 1, ..., m_i$, m_i is the number of urine measures for subject i, and $\theta'_0, \theta'_1, \theta'_2$ are the regression parameters. Estimations of $\theta'_0, \theta'_1, \theta'_2$, i.e., $\hat{\theta}'_0, \hat{\theta}'_1, \hat{\theta}'_2$ are used to model contamination probability $\hat{p}_{it} = \hat{P}r(R_{it} = 1)$.

Hence, $\hat{p}_{it} = \hat{P}r\left(R_{it}=1|X_i, B_{it}, \hat{\theta}'\right)$ is estimated by the following model:

$$\log \frac{\hat{P}r\left(R_{it}=1|X_i, B_{it}, \hat{\theta}'\right)}{1-\hat{P}r\left(R_{it}=1|X_i, B_{it}, \hat{\theta}'\right)} = \hat{\theta}'_0 + \hat{\theta}'_1 X_i + \hat{\theta}'_2 B_{it},$$

$$\hat{p}_{it} = \frac{e^{\hat{\theta}'_0 + \hat{\theta}'_1 X_i + \hat{\theta}'_2 B_{it}}}{1 + e^{\hat{\theta}'_0 + \hat{\theta}'_1 X_i + \hat{\theta}'_2 B_{it}}},$$

where $i = 1, ..., N; t = 1, ..., T$.

If the mean model and the contamination probability for each observation are cor-

rectly specified, this MCGEE method may provide a working estimate of regression parameters under certain assumptions. However, challenges remain in an accurate estimation of p_{it}. In other words, the estimate of β may not be consistent when the contamination probability is misspecified. In the next two sections, we address the asymptotic normality of estimators from MCGEE under the true value of contamination probability, and examine the asymptotic bias of estimators based on the MCGEE for self-reported data when the estimation of the contamination probability is biased.

2.3 Asymptotic Properties of the Estimators

For the true drug use data Y_{it}, where $i = 1, ..., N, t = 1, ..., T$, the asymptotic properties have been derived by Liang and Zeger(1986). Denoting $\mu_i^* = E(Z_i|X_i, \beta) = (\mu_{i1}^*, ...\mu_{iT}^*)'$, the GEE for Y_i are:

$$U_\beta(\beta) = \sum_{i=1}^{N} D_i' V_i^{-1} (Y_i - \mu_i) = 0,$$

where $D_i = \partial\mu_i/\partial\beta$, and V_i is the covariance matrix of Y_i. V_i can be decomposed into the form $A_i^{\frac{1}{2}} C_i(\gamma) A_i^{\frac{1}{2}}$, where A_i is a matrix with the marginal variances on the main diagonal and zeros elsewhere, γ is a vector which fully characterize $C_i(\gamma)$, and $C_i(\gamma)$ is a working correlation matrix of Y_i's.

Under the assumption that the estimator for correlation parameter γ is $\sqrt{N}$ consistent given β, Liang and Zeger (1986) showed that $\sqrt{N}(\hat{\beta} - \beta)$ is asymptotically

normally distributed with mean zero and estimated variance matrix

$$\left(\sum_{i=1}^{N}\hat{D}_i'\hat{V}_i^{-1}\hat{D}_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{D}_i'\hat{V}_i^{-1}\hat{A}_i\hat{V}_i^{-1}\hat{D}_i\right)\left(\sum_{i=1}^{N}\hat{D}_i'\hat{V}_i^{-1}\hat{D}_i\right)^{-1},$$

and the result does not depend on the choice of the correlation matrix $C_i(\gamma)$.

In this section, we study the asymptotic properties of $\hat{\beta}$, the solution of the MCGEE of self-reported data, by proving its existence, its consistency and its asymptotic normality as sample size $N \to \infty$ and taking the time periods for each subject, T, to be bounded for all subjects. We build our study upon the seminal work of Liang and Zeger (1986), and Yuan and Jennrich (1998). However, our research differs from those two by considering an additional property in MCGEE, the contamination probability of self-reported binary outcomes.

The MCGEE form of the self-reported data Z_i is:

$$\begin{aligned} U_N(\beta) &= \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1}(Z_i - \mu_i^*) \\ &= \sum_{i=1}^{N}(1-2p_i)\otimes X_i' A_i (A_i^{*\frac{1}{2}} C_i^*(\gamma) A_i^{*\frac{1}{2}})^{-1}(Z_i - (\mu_i - 2\mu_i \otimes p_i + p_i)), \end{aligned}$$

where the correlation parameter γ and the contamination probabilities p_i are assumed to be known.

Here we aim to show that the solution $\hat{\beta}_N$ to $U_N(\beta) = 0$ is consistent,

$$\hat{\beta}_N \to \beta_0,$$

almost surely for the true value β_0, and $\hat{\beta}_N$ is approximately normally distributed as $N \to \infty$.

The Assumptions we need to prove the consistency and asymptotic normality of $\hat{\beta}_N$ are:

Assumption A. The subjects are independently sampled and there exists an upper bound $M < \infty$ such that the number of replicates $m_i < M$ for all subjects $i = 1, 2, ...$

Assumption B. There exists an upper bound $b < \infty$ such that $|X_i| < b$ for all subjects $i = 1, 2, ...$

Assumption C. It is assumed that $\frac{1}{N}\sum_{i=1}^{N} X_i X_i' \to B$ as $N \to \infty$, where B is a positive definite matrix.

Assumption A ensures that data from a finite number of subjects and do not dominate the parameter estimator. Assumption B ensures that the estimating functions $\frac{1}{N}U_N(\beta)$ and its first and higher-order derivatives with respect to beta are bounded. Assumption C means that for sufficiently large N, $\frac{1}{N}E(\frac{\partial}{\partial\beta'}U_N(\beta))$ will be positive definite, and there is no redundancy in the predictors.

Define the matrices

$$I_0^*(\beta) = \lim_{N\to\infty} \frac{1}{N} \frac{\partial}{\partial \beta'} U_N(\beta) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (D_i^{*\prime} V_i^{*-1} D_i^*),$$

and

$$I_1^*(\beta) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (D_i^{*\prime} V_i^{*-1} A_i^* V_i^{*-1} D_i^*).$$

The existence of these limits is ensured by Assumption B. Moreover, Assumption C ensures that $I_0^*(\beta)$ and $I_1^*(\beta)$ are positive definite.

The following Theorem shows that the MCGEE is strongly consistent for β.

Theorem 1. *Under Assumptions A-C, with probability one there exist zeros* $\hat{\beta}_N$ *of* $U_N(\beta) = 0$ *such that* $\hat{\beta}_N \to \beta_0$ *as* $N \to \infty$.

In the Appendix, we demonstrate that $\frac{1}{N} U_N(\beta_0) \to 0$ a.s., as $N \to \infty$, and that $\frac{1}{N} \frac{\partial}{\partial \beta'} U_N(\beta)$ converges uniformly to a non-stochastic limit which is nonsingular at β_0. The results then follow from Theorem 2 of Yuan and Jennrich (1997).

The following Theorem shows that the MCGEE estimator is approximately normally distributed for large N.

Theorem 2. *Under Assummptions A-C,* $\sqrt{N}\left(\hat{\beta} - \beta_0\right) \underset{\sim}{L} N\left(0, I_0^{*-1}(\beta_0) I_1^*(\beta_0) I_0^{*-1}(\beta_0)\right)$, *as* $N \to \infty$.

By Theorem 4 of Yuan and Jennrich (1998), it suffices to prove that

$$\frac{1}{\sqrt{N}} U_N(\beta_0) \underrightarrow{L} N(0, I_1^*(\beta_0)),$$

as $N \to \infty$, which is proved in the appendix.

The variance-covariance estimator of $\hat{\beta}$ can be estimated as:

$$\hat{I}_0^* = \sum_{i=1}^{N} (\hat{D}_i^{*\prime} \hat{V}_i^{*-1} \hat{D}_i^*),$$

and

$$\hat{I}_1^* = \sum_{i=1}^{N} \hat{D}_i^{*\prime} \hat{V}_i^{*-1} \hat{A}_i^* \hat{V}_i^{*-1} \hat{D}_i^*.$$

Therefore, the asymptotic properties of $\hat{\beta}$ holds with the accurate estimation of the contamination probability. These asymptotic properties also hold under consistent estimates of the correlation parameter γ and contamination probabilities p_i. However, if the estimation of this probability is misspecified, $\hat{\beta}$ may not be asymptotically unbiased. In the next section, we examine the asymptotic bias of $\hat{\beta}$ from the MCGEE approach for self-reported data when the estimation of the contamination probability deviated from the true value.

2.4 Bias of the Estimators with Report Bias

The main concern we have in estimating β from the MCGEE approach of self-reported data is having biased estimation of the contamination probability p_{it}. In this section, we estimate the asymptotic bias of the MCGEE estimator $\hat{\beta}$.

The MCGEE form of the self-reported data Z_i is:

$$\begin{aligned} U^*_\beta(\beta) &= \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} (Z_i - \mu_i^*) \\ &= \sum_{i=1}^{N} (1 - 2p_i) \otimes X_i' A_i (A_i^{*\frac{1}{2}} C_i^*(\gamma) A_i^{*\frac{1}{2}})^{-1} (Z_i - (\mu_i - 2\mu_i \otimes p_i + p_i)) . \end{aligned}$$

From section 2.2, the contamination probability p_i can be estimated using two methods.

One method is assuming each subject has one single $E(R_{it})$ for all time points, then the contamination probability $\hat{p}_i$ for the ith subject can be estimated as:

$$\hat{p}_i = \frac{\sum_{j=1}^{m} F_{ij}}{m_i},$$

where $i = 1, ..., N; j = 1, ..., m_i$, m_i is the number of urine measures for subject i, F_{ij} is the difference between urine results and the indicator of the summation of self-reported cocaine use of h days.

The other method is to assume that the contamination indicator R_{it} depends on

some covariates and can be modeled through logistic regression models. $\hat{p}_{it}$ can be estimated as:

$$\hat{p}_{it} = \frac{e^{\hat{\theta}_0' + \hat{\theta}_1' X_i + \hat{\theta}_2' B_{it}}}{1 + e^{\hat{\theta}_0' + \hat{\theta}_1' X_i + \hat{\theta}_2' B_{it}}},$$

where $i = 1, ..., N; t = 1, ..., T$, X_i denotes a vector of time independent covariate; B_{it} denotes a vector of time dependent covariate.

Assume the marginal mean of the true drug use for the ith subject is not equal to 0.5 for each time points, $\mu_i = (\mu_{i1}, ...\mu_{iT})' \neq (0.5, ..., 0.5)'$, the contamination probability is not equal to 0.5 for each time points, i.e., $p_i = (p_{i1}, ...p_{iT})' \neq (0.5, ..., 0.5)'$, and the estimated contamination probability is also not equal to 0.5 for every time points, $\hat{p}_i = (p_{i1}, ...p_{iT})' \neq (0.5, ..., 0.5)'$.

If we replace p_i by $\hat{p}_i$ in the estimating equation, then $E_{\beta_0}(U_\beta^*(\beta))$ may not be equal to 0.

$$\begin{aligned} E_{\beta_0}(U_\beta^*(\beta)) &= \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} \left(E(Z_i) - \mu_i^*\right) \\ &= \sum_{i=1}^{N} (1 - 2\hat{p}_i) X_i' A_i (A_i^{*\frac{1}{2}} C_i^*(\gamma) A_i^{*\frac{1}{2}})^{-1} (1 - 2\mu_i)(p_i - \hat{p}_i). \end{aligned}$$

The above equation only equals 0 when $\hat{p}_i = p_i$, which means when the contamination probability has been correctly estimated, we can have unbiased estimators.

From section 2.3, by **Theorem 1** we have

$$E_{\beta_0}(U^*_\beta(\beta)) = \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} \left(E(Z_i) - \mu_i^*\right) = 0,$$

$$\hat{\beta}_N \to \beta_0,$$

as $N \to \infty$, where β_0 is the true value.

However, with $\hat{p}_i \neq p_i$, $(E(Z_i) - \mu_i^*) \neq 0$, there exists bias of $\hat{\beta}$,

$$E_{\beta_0}(U^*_\beta(\beta)) = \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} \left(E(Z_i) - \mu_i^*\right) \neq 0.$$

Instead, we can only have

$$E_{\beta_0}(U^*_\beta(\beta^*)) = 0,$$

$$\hat{\beta}_N \to \beta^*,$$

as $N \to \infty$.

To estimate the asymptotic bias of $\hat{\beta}$ is equivalent to calculate $\beta^* - \beta_0$. Intuitively, when the difference between $\hat{p}_i$ and p_i increases, i.e., the difference between $E(Z_i)$ and μ_i^* increases, the bias of $\hat{\beta}$ increases subsequently. Unfortunately, the above equation does not have a closed form solution of β. Thus, we may borrow the idea

from Rotnitzky and Wypij (1992), since for any fixed β, the estimating equation is a function of (Z_i, R_i, X_i). It has expectation given by the sum of all the possible situations times their respected probabilities. Then instead of solving β from the above equations, we may simply consider an artificial sample comprised of one observation for each possible combinations of (Z_i, R_i, X_i), which weighted by their specific probabilities.

2.5 Simulations

2.5.1 Data generation

Two groups (treatment and control) are considered in the data generation. In each group, there are $N/2$ subjects whose outcomes are repeatedly measured at T time points. True drug use data is generated as:

$$Pr\left(Y_{it} = 1 | X_i, \beta\right) = \mu_{it},$$

$$\log \frac{\mu_{it}}{1 - \mu_{it}} = X_i\beta = \beta_0 + \beta_1 X_i + \sigma_i,$$

for $i = 1, ..., N$, $t = 1, ..., T$, where X_i is the treatment indicator, $X_i = 1$ denotes the individual in the treatment group, $X_i = 0$ denotes the individual in the control group, σ_i is a random effect variable following a normal distribution with mean zero and a common variance $v = 0.04$.

Urine data is generated based on the true drug use data. For the ith patient, we use U_{ij} to denote urine test result, where $j = 1, ..., m_i$. m_i is the number of urine measurements for subject i. Based on the notation used previously, the self-reported data can be written as:

$$Z_{it} = Y_{it}\left(1 - R_{it}\right) + \left(1 - Y_{it}\right) R_{it}.$$

Contamination indicator R_{it} is generated using two methods. First, assuming each subject has one single $E(R_{it})$ for all time points, R_{it} is generated by a relatively simple contamination probability assumption, which is not model based. Assuming there is p_1 probability of contamination among all subjects at one or several time points, and within these subjects, there is p_2 probability that they report false drug use at each time point. Each observation is independent. Thus, the overall contamination probability p equals:

$$p = p_1 \times p_2.$$

And it can be estimated as:

$$\hat{p} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} R_{it}}{N \times T}.$$

For each contaminated subjects, the contamination probability at each time point

can be estimated as:

$$\hat{p}_i = \frac{1}{T}\sum_{t=1}^{T} R_{it},$$

the probability of how many subjects have been contaminated has the estimated form:

$$\hat{p}_1 = \frac{1}{N}\sum_{i=1}^{N} I_i \left(\sum_{t=1}^{T} R_{it} \geq 1\right),$$

where I_i is the indicator, $I_i = 1$ if $\sum_{t=1}^{T} R_{it} \geq 1$; $I_i = 0$, otherwise.

Second, assuming the contamination indicator R_{it} depends on some covariates, it is generated by a model based on time dependent covariates B_{it} and time independent covariates X_i. For instance, X_i can be the treatment effect, B_{it} can be the buprenorphone bottle open data, σ_i is a random effect variable following a normal distribution with mean zero and a common variance $v = 0.04$. In our simulation, we assume the indicator variable R_{it} follows:

$$\log \frac{Pr\left(R_{it} = 1 | X_i, B_{it}, \theta\right)}{1 - Pr\left(R_{it} = 1 | X_i, B_{it}, \theta\right)} = \theta_0 + \theta_1 X_i + \theta_2 B_{it} + \sigma_i.$$

Bias of estimators corresponding to each generation method have been assessed under several different scenarios of contamination probabilities, different time periods for cocaine to be cleared from urine, and sample sizes.

All data simulations and analysis are carried out using the R software, 1000 replications are performed for each run to obtain reliable results.

2.5.2 Simulation results

In this section, we analyze the bias and standard error of the estimators of the GEE approach and the Mean Corrected GEE (MCGEE) approach for various situations under simulation. We assume that urine samples are collected every 7 days ($k = 7$). We generate $N/2$ subjects in the treatment and the control group respectively whose outcomes are repeatedly measured at T time points. The true value for the intercept β_0 is 0.3, and the true value for the treatment effect β_1 is 1.2.

First, we consider the case that R_{it} is generated by a relatively simple contamination probability assumption. As we discussed earlier, the bias of the MCGEE approach of self-reported data exist when the estimated contamination probability is not the same as the true value ($\hat{p} \neq p$). Several combinations for different sample size ($N = 100, N = 400$), time points for each individual's measurement ($T = 70, T = 140$), contamination probabilities ($p_1 = 0.4, 0.6; p_2 = 0.4, 0.8$), and different time periods for cocaine to be cleared from urine ($h = 1, h = 4$) have been studied.

Table 2.1 reports the parameters' estimation, standard error, and the coverage probability of 95% confidence intervals of the GEE approach and the MCGEE approach

for different combinations of sample size, time points, and contamination probability under the assumption that the time period for cocaine to be cleared from urine is 1 day. In this situation, the difference of the estimated contamination probability and the true contamination probability is relatively small. We can find out that when p_1 is fixed, as p_2 increases, both the intercept's and the treatment effect's bias of the GEE approach increase. Similarly, when p_2 is fixed, as p_1 increases, both these two estimators' bias of the GEE approach increase. This pattern can be observed for the MCGEE approach as well, but for each p_1 and p_2's combination, the bias of the MCGEE estimators are much lower than the bias of the GEE estimators. The difference of biases of these two approaches are highly dependent on the contamination probability. When $p_1 = p_2 = 0.4$, the difference of parameters' estimation of GEE and MCGEE is relatively small. When p_1 or p_2 increases, this difference increases significantly. When $p_1 = 0.6$ and $p_2 = 0.8$, the difference of parameters' estimation of GEE and MCGEE is large. The bias correction effect of the MCGEE approach can be easily noticed under higher contamination probability. The results of the coverage probability of 95% confidence intervals suggest similar findings. Table 2.1 also indicates that the standard error of parameters of both methods decrease as sample size and time periods increase. The results of bias of the treatment estimator of GEE and MCGEE approach are also presented in Figure 2.2.

After increasing the time period for cocaine to be cleared from urine to 4 days ($k = 7, h = 4$), we observe that the bias of the parameter estimates of the MCGEE approach are still lower than the GEE approach. However, the performances of the

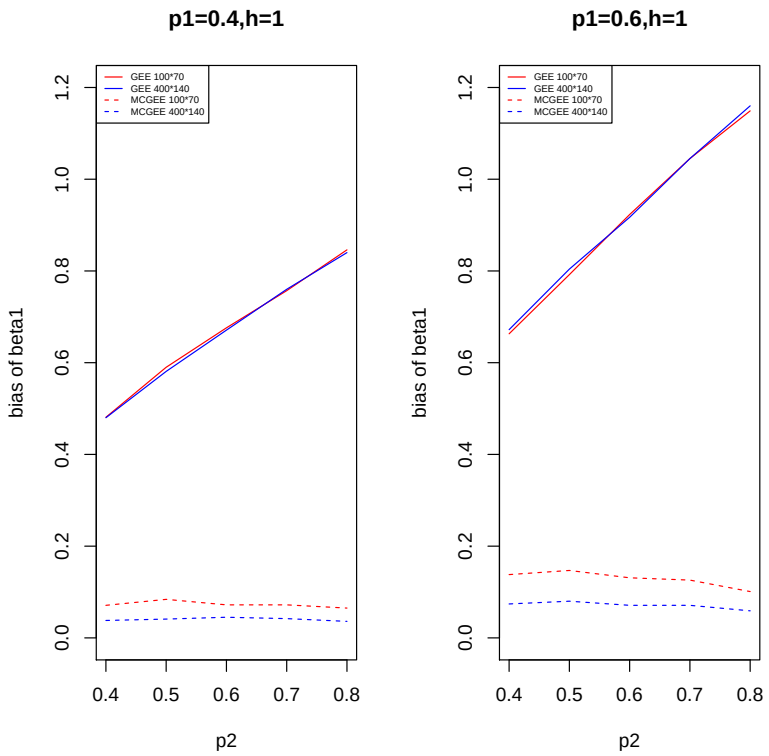
Table 2.1: Parameters' estimation, standard error (S.E.), and the coverage probability of 95% CI (CP%) of the GEE approach, and the MCGEE approach (k=7, h=1).

Effect	N	T	p_1	p_2	GEE	CP%	MCGEE	CP%
$\beta_0(S.E.)$	100	70	0.4	0.4	0.200(0.04)	36.4	0.279(0.05)	94.0
				0.8	0.111(0.05)	5.4	0.291(0.05)	94.6
			0.6	0.4	0.153(0.04)	5.6	0.262(0.06)	91.8
				0.8	0.010(0.05)	0.0	0.283(0.06)	94.2
		140	0.4	0.4	0.202(0.03)	26.4	0.292(0.05)	94.6
				0.8	0.107(0.04)	2.2	0.293(0.04)	94.6
			0.6	0.4	0.152(0.03)	1.4	0.274(0.05)	92.4
				0.8	0.016(0.05)	0.0	0.292(0.04)	95.0
	400	70	0.4	0.4	0.202(0.02)	0.8	0.283(0.03)	90.4
				0.8	0.105(0.02)	0.0	0.289(0.02)	92.0
			0.6	0.4	0.153(0.02)	0.0	0.267(0.03)	82.6
				0.8	0.012(0.02)	0.0	0.282(0.03)	89.8
		140	0.4	0.4	0.200(0.02)	0.0	0.288(0.02)	91.2
				0.8	0.106(0.02)	0.0	0.293(0.02)	92.2
			0.6	0.4	0.154(0.02)	0.0	0.281(0.03)	89.0
				0.8	0.012(0.02)	0.0	0.290(0.02)	92.3
$\beta_1(S.E.)$	100	70	0.4	0.4	0.719(0.10)	0.2	1.129(0.09)	87.6
				0.8	0.354(0.16)	0.0	1.135(0.08)	86.4
			0.6	0.4	0.537(0.09)	0.0	1.062(0.11)	74.2
				0.8	0.051(0.15)	0.0	1.099(0.09)	79.0
		140	0.4	0.4	0.719(0.09)	0.0	1.162(0.07)	92.0
				0.8	0.350(0.16)	0.0	1.161(0.07)	91.2
			0.6	0.4	0.535(0.08)	0.0	1.131(0.09)	88.6
				0.8	0.033(0.14)	0.0	1.141(0.07)	88.0
	400	70	0.4	0.4	0.713(0.05)	0.0	1.128(0.05)	65.4
				0.8	0.360(0.08)	0.0	1.137(0.04)	68.0
			0.6	0.4	0.529(0.05)	0.0	1.055(0.06)	25.4
				0.8	0.041(0.07)	0.0	1.099(0.04)	35.6
		140	0.4	0.4	0.720(0.05)	0.0	1.162(0.04)	79.6
				0.8	0.360(0.06)	0.0	1.164(0.03)	82.6
			0.6	0.4	0.528(0.04)	0.0	1.126(0.04)	59.4
				0.8	0.040(0.07)	0.0	1.141(0.04)	62.5

Table 2.2: Parameters' estimation standard error (S.E.), and the coverage probability of 95% CI (CP%) of the GEE approach, and the MCGEE approach (k=7, h=4).

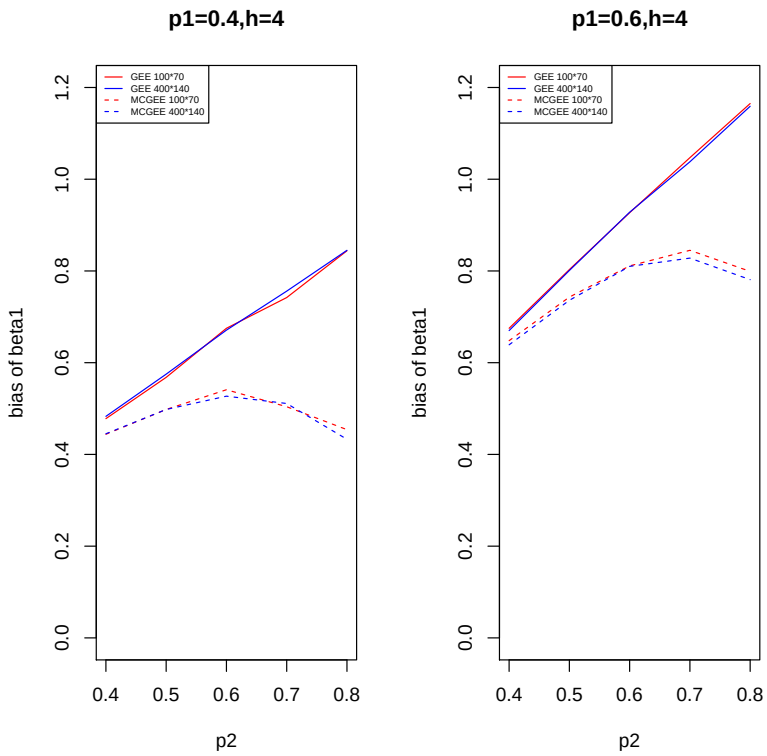
Effect	N	T	p_1	p_2	GEE	CP%	MCGEE	CP%
$\beta_0(S.E.)$	100	70	0.4	0.4	0.199(0.04)	34.8	0.225(0.04)	58.4
				0.8	0.109(0.05)	5.6	0.157(0.06)	26.2
			0.6	0.4	0.155(0.04)	7.2	0.188(0.04)	30.0
				0.8	0.010(0.05)	0.2	0.058(0.06)	2.2
		140	0.4	0.4	0.201(0.03)	22.6	0.225(0.04)	47.6
				0.8	0.105(0.04)	2.2	0.150(0.05)	13.0
			0.6	0.4	0.153(0.03)	1.6	0.183(0.04)	10.2
				0.8	0.014(0.04)	0.0	0.058(0.05)	0.6
	400	70	0.4	0.4	0.201(0.02)	0.2	0.228(0.02)	8.6
				0.8	0.108(0.03)	0.0	0.156(0.03)	0.0
			0.6	0.4	0.154(0.02)	0.0	0.187(0.02)	0.0
				0.8	0.010(0.03)	0.0	0.059(0.03)	0.0
		140	0.4	0.4	0.201(0.02)	0.2	0.227(0.02)	4.6
				0.8	0.107(0.02)	0.0	0.153(0.03)	0.0
			0.6	0.4	0.153(0.01)	0.0	0.183(0.02)	0.0
				0.8	0.013(0.02)	0.0	0.057(0.03)	0.0
$\beta_1(S.E.)$	100	70	0.4	0.4	0.722(0.10)	0.8	0.756(0.10)	1.6
				0.8	0.356(0.15)	0.0	0.746(0.16)	16.0
			0.6	0.4	0.525(0.10)	0.0	0.552(0.10)	0.0
				0.8	0.035(0.15)	0.0	0.401(0.19)	2.6
		140	0.4	0.4	0.721(0.09)	0.4	0.759(0.09)	0.6
				0.8	0.377(0.15)	0.0	0.782(0.15)	20.0
			0.6	0.4	0.536(0.09)	0.0	0.566(0.09)	0.0
				0.8	0.032(0.14)	0.0	0.406(0.19)	1.4
	400	70	0.4	0.4	0.715(0.05)	0.0	0.750(0.05)	0.0
				0.8	0.349(0.08)	0.0	0.743(0.08)	0.0
			0.6	0.4	0.527(0.04)	0.0	0.554(0.05)	0.0
				0.8	0.041(0.07)	0.0	0.412(0.09)	0.0
		140	0.4	0.4	0.717(0.04)	0.0	0.755(0.05)	0.0
				0.8	0.355(0.07)	0.0	0.767(0.07)	0.0
			0.6	0.4	0.530(0.04)	0.0	0.561(0.05)	0.0
				0.8	0.041(0.07)	0.0	0.419(0.09)	0.0

Figure 2.2: Bias of β_1 of GEE and MCGEE when $h = 1$



MCGEE approach on bias correction decline significantly (Table 2.2, Figure 2.3). The main reason is the bias of the contamination probability estimation $\hat{p}$ increases as the time period for cocaine to be cleared from urine increases. For instance, even if we detect a contamination in a certain time block, because of the extended period of cocaine clearance from urine, it's still difficult to locate the exact time point for that contamination. When p_2 is high, e.g., $p_2 = 0.8$, which indicates that there is 80% contamination probability within each contaminated subject, the difference of

Figure 2.3: Bias of β_1 of GEE and MCGEE when $h = 4$



parameters' estimation between GEE and MCGEE is larger, as it's easier to detect the contamination and correct the marginal mean of the GEE approach.

Second, we consider the case that R_{it} is generated by a model based on the bottle open data B_{it} and the treatment indicator X_i:

$$\log \frac{Pr\left(R_{it} = 1 | X_i, B_{it}, \theta\right)}{1 - Pr\left(R_{it} = 1 | X_i, B_{it}, \theta\right)} = \theta_0 + \theta_1 X_i + \theta_2 B_{it} + \sigma_i.$$

Table 2.3: Parameter value for model based generation of R_{it}.

$\bar{p}$	θ_0	θ_1	θ_2
0.1	0.05	0.5	-8.0
0.3	0.8	4.5	-3.0
0.5	1.4	5.5	-2.0

Several contamination probabilities estimated by the true R_{it} have been evaluated. The θ values we used to generate R_{it} for different mean of contamination probabilities are in Table 2.3.

From Table 2.4, 2.5 and Figure 2.4, we observe that: 1. when contamination probability p increases, the bias of both the GEE estimators and MCGEE estimators increase; 2. for each situation, the bias of the MCGEE estimators are much lower than the bias of the GEE estimators; 3. when the time period for cocaine to be cleared from urine increases from 1 day to 4 days, the bias of both the GEE and MCGEE approaches increase; 4. the differences between the bias of the MCGEE and the GEE estimators drop significantly when the time period for cocaine to be cleared from urine increases; 5. when sample size N and time periods T increase, the standard error of both the GEE and MCGEE estimators decreases.

Similar as the previous situation, when the time period for cocaine to be cleared from urine increases from 1 day to 4 days, it's difficult to estimate the true contamination probability, hence the marginal mean of the GEE model. The performance of the MCGEE approach significantly associates with the time period for cocaine to

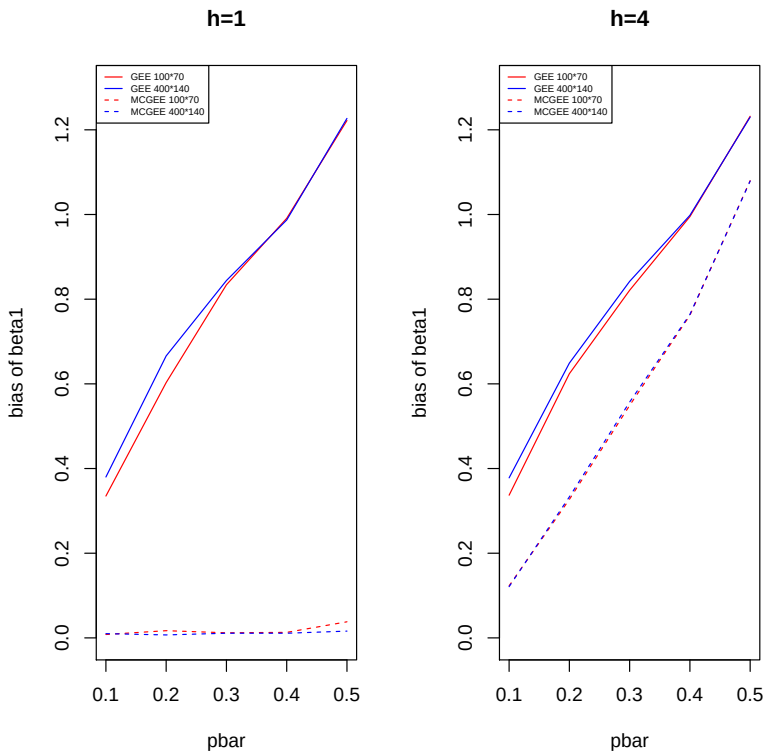
Table 2.4: Parameters' estimation standard error (S.E.), and the coverage probability of 95% CI (CP%) of the GEE approach, and the MCGEE approach (k=7, h=1).

Effect	N	T	$\bar{p}$	GEE	CP(%)	MCGEE	CP(%)
$\beta_0(S.E.)$	100	70	0.1	0.234(0.04)	65.4	0.293(0.04)	85.8
			0.3	0.110(0.05)	3.2	0.293(0.04)	88.6
			0.5	-0.008(0.04)	0.0	0.298(0.06)	90.2
		140	0.1	0.229(0.04)	54.4	0.299(0.03)	82.6
			0.3	0.102(0.04)	0.4	0.294(0.03)	84.0
			0.5	-0.010(0.04)	0.0	0.296(0.04)	83.4
	400	70	0.1	0.236(0.02)	18.4	0.299(0.02)	89.0
			0.3	0.110(0.02)	0.0	0.298(0.02)	89.0
			0.5	-0.007(0.02)	0.0	0.299(0.03)	89.6
		140	0.1	0.225(0.02)	3.4	0.297(0.01)	78.0
			0.3	0.106(0.02)	0.0	0.298(0.02)	84.4
			0.5	-0.009(0.02)	0.0	0.297(0.02)	85.2
$\beta_1(S.E.)$	100	70	0.1	0.865(0.10)	10.8	1.192(0.09)	95.4
			0.3	0.366(0.15)	0.0	1.188(0.13)	91.6
			0.5	-0.022(0.11)	0.0	1.162(0.10)	90.6
		140	0.1	0.824(0.11)	6.2	1.187(0.08)	94.2
			0.3	0.363(0.15)	0.0	1.192(0.10)	93.8
			0.5	-0.023(0.10)	0.0	1.180(0.11)	76.8
	400	70	0.1	0.865(0.05)	0.0	1.188(0.04)	94.8
			0.3	0.364(0.07)	0.0	1.187(0.06)	88.2
			0.5	-0.029(0.05)	0.0	1.178(0.05)	87.0
		140	0.1	0.820(0.05)	0.0	1.190(0.05)	94.8
			0.3	0.356(0.06)	0.0	1.189(0.04)	91.8
			0.5	-0.027(0.05)	0.0	1.184(0.04)	78.6

Table 2.5: Parameters' estimation standard error (S.E.), and the coverage probability of 95% CI (CP%) of the GEE approach, and the MCGEE approach(k=7, h=4).

Effect	N	T	$\bar{p}$	GEE	CP(%)	MCGEE	CP(%)
$\beta_0(S.E.)$	100	70	0.1	0.238(0.04)	68.0	0.274(0.04)	79.4
			0.3	0.107(0.05)	2.6	0.177(0.04)	17.8
			0.5	-0.005(0.04)	0.0	0.039(0.04)	0.0
		140	0.1	0.226(0.03)	49.0	0.271(0.03)	67.8
			0.3	0.110(0.04)	0.8	0.182(0.03)	6.8
			0.5	-0.010(0.04)	0.0	0.034(0.03)	0.0
	400	70	0.1	0.236(0.02)	19.2	0.273(0.02)	63.6
			0.3	0.112(0.02)	0.0	0.183(0.02)	0.0
			0.5	-0.007(0.02)	0.0	0.037(0.02)	0.0
		140	0.1	0.224(0.02)	1.4	0.269(0.01)	37.8
			0.3	0.106(0.02)	0.0	0.179(0.02)	0.0
			0.5	-0.009(0.02)	0.0	0.036(0.02)	0.0
$\beta_1(S.E.)$	100	70	0.1	0.863(0.10)	9.0	1.077(0.09)	80.4
			0.3	0.379(0.15)	0.0	0.651(0.11)	0.2
			0.5	-0.032(0.11)	0.0	0.120(0.10)	0.0
		140	0.1	0.825(0.11)	6.0	1.075(0.09)	85.4
			0.3	0.348(0.15)	0.0	0.637(0.08)	0.0
			0.5	-0.031(0.10)	0.0	0.124(0.08)	0.0
	400	70	0.1	0.866(0.05)	0.0	1.082(0.06)	39.6
			0.3	0.368(0.08)	0.0	0.648(0.06)	0.0
			0.5	-0.027(0.05)	0.0	0.122(0.04)	0.0
		140	0.1	0.822(0.05)	0.0	1.079(0.04)	34.6
			0.3	0.358(0.07)	0.0	0.643(0.04)	0.0
			0.5	-0.030(0.05)	0.0	0.121(0.04)	0.0

Figure 2.4: Bias of β_1 of GEE and MCGEE



be cleared from urine.

2.6 SCU Data

In the Self-reported Cocaine use and Urine test (SCU) data, there are a total of 140 patients, followed for a period of 5-6 months. All enrolled patients had met the current DSM-IV diagnosis criteria for cocaine dependence (Argoff and McCleane 2009).

After a 2-week induction and stabilization period, during which patients were treated by nurses 3 times per week with 16 mg of buprenorphine daily, enrolled subjects were randomly assigned to the treatment or the control group. Both groups received buprenorphine, a substitute for cocaine use, which was stored in bottles. The special MEMSCAP bottles can record the time when the bottle is opened. Buprenorphine was instructed to be taken once per day. If the bottle was opened on a specific day, the patient was regarded as adherent for that day.

The control group received physical management (PM), a 15-20 minutes session by Internal Medicine physicians with experience as buprenorphine providers. Throughout the study period, sessions occurred weekly for the first two weeks, every two weeks for the next four weeks and then monthly. The treatment group received PM plus cognitive behavioral therapy (CBT). CBT is a counseling intervention that has demonstrated efficacy in treating a variety of psychiatric conditions and cocaine dependences. CBT was provided by masters- and doctoral-level clinicians who were trained with a manual adapted from a guidance for the use of CBT for cocaine dependence (Carroll 1998). The main components of counseling focused on performing a functional analysis of behavior, promoting behavioral activation, identifying and coping with drug cravings, enhancing drug refusal skills and decision makings about high risk situations, and improving problem solving skills (Fiellin et al. 2013).

The study's major outcomes include (1) self-reported daily illicit drug uses which were reported during the weekly PM sessions, and (2) weekly urine test results. Self-

reported daily illicit drug uses variables include cocaine use, marijuana use, alcohol use, bup use, and prescopioid/heroin/opium/other opiate use. Urine test variables include cocaine, benzo, THC, and opiate/methadone/oxycontin. Overall, there are five self-reported variables and four urine test variables. Our main interest in the motivating example is the cocaine use, including the self-reported cocaine use and the weekly urine test on cocaine. After statistical methods have been developed for the SCU data on these two variables, they may also be extended to other substances' uses.

As defined in the previous sections, let Z_{it} denote the self-reported daily cocaine use, X_i denote the treatment effect, with $X_i = 1$ indicating patient in the treatment group. We conduct the GEE model (independent correlation matrix) using self-reported data to simply replace true data with the logistic link:

$$\log \frac{Pr\left(Z_{it} = 1 | X_i, \beta\right)}{1 - Pr\left(Z_{it} = 1 | X_i, \beta\right)} = \beta_0 + \beta_1 X_i.$$

Table 2.6: Results of GEE analysis of cocaine use

	Estimate	S.E.	P-value
β_0	-2.15	0.11	<0.0001
β_1	0.23	0.16	0.15

The results of the GEE analysis is presented in Table 2.6. The p-value for effect of treatment is 0.15, indicating that there is no significant treatment effect.

However, Z_{it} is potentially contaminated, which may lead to bias. A further analysis

comparing self-reported cocaine use and urine test results at each time point urine was collected showed that the contamination probability is around 21.4% (Table 2.7).

Table 2.7: Results of self-reported cocaine use and urine test results

	Urine test negative	Urine test positive
Self-reported negative	1500	411
Self-reported positive	2	19

Table 2.8: Results of MCGEE approach of cocaine use (h=1)

	Estimate	S.E.	P-value
β_0	-2.71	0.11	<0.0001
β_1	0.29	0.16	0.07

Thus, we use urine test results to estimate the contamination probability and the mean of self-reported data. We then apply the MCGEE approach to estimate the effect of CBT.

First, we assume the contamination indicator R_{it} does not depend on the MEMSCAP bottle open data, and the time period for cocaine to be cleared from urine is 1 day ($h = 1$). From the above table, we can clearly see that the estimation of treatment effect has been increased, the p-value for the effect of treatment is 0.07. (Table 2.8).

We now consider the case when $h = 4$, from Table 2.9, the estimation of treatment effect increases to 0.30 comparing with the situation where $h = 1$, and the p-value drops to 0.03, indicating that there is a significant treatment effect.

Table 2.9: Results of MCGEE approach of cocaine use (h=4)

	Estimate	S.E.	P-value
β_0	-2.47	0.10	<0.0001
β_1	0.30	0.14	0.03

Second, we assume the contamination indicator R_{it} depends on the MEMSCAP bottle open indicator. Tables 2.10 and 2.11 show the result for the situation $h = 1$, and $h = 4$. The effects of treatment are similar for these two cases, and both p-values are less than 0.05, suggesting these effects are statistically significant.

Table 2.10: Results of MCGEE approach of cocaine use (h=1)

	Estimate	S.E.	P-value
β_0	-2.31	0.09	<0.0001
β_1	0.27	0.13	0.04

Table 2.11: Results of MCGEE approach of cocaine use (h=4)

	Estimate	S.E.	P-value
β_0	-2.38	0.09	<0.0001
β_1	0.27	0.13	0.04

2.7 Discussion and Conclusion

In this chapter, we proposed the Mean Corrected GEE (MCGEE) approach for analyzing the longitudinal binary self-reported outcomes with report bias. When the marginal mean of the self-reported results Z_{it} has been correctly specified, this

MCGEE approach yields consistent estimates of the parameters. These estimates are consistent whether or not the working correlation matrix is specified. Urine test results are used to estimate the contamination probability and correct the marginal means of the self-reported results. However, urine test is measured weekly, and the time period for cocaine to be cleared from urine can be longer than one day. From our study, we noticed that it is challenging to accurately estimate the contamination probability when the time period for cocaine to be cleared from urine is as long as 4 days. Therefore, a limitation of our proposed approach is that consistency of the estimators of the model parameters requires the correct estimation of the contamination probability. In other words, estimators may not be asymptotically unbiased when the contamination probability is misspecified, i.e., the marginal means have not been correctly estimated.

Comparing the GEE approach, which simply replaced the true data Y_{it} with self-reported data Z_{it}, with the MCGEE approach, we find that in every situation when we combine different sample sizes, contamination probabilities, and the time periods for cocaine to be cleared from urine, the bias of the MCGEE estimators are lower than the bias of the GEE estimators. The bias of the parameters' estimation of the GEE approach increases as contamination probability increase for both contamination indicator assumptions. When the time period for cocaine to be cleared from urine is 1 day ($h = 1$), we can successfully detect the contamination in the self-reported data for most cases, the bias of the parameters' estimation of the MCGEE approach is relatively small, and the coverage probability of 95% CI of MCGEE is

around 95%. And the bias of the parameters' estimation of the MCGEE approach remains similar when contamination probability increases. These findings suggest that we can consistently detect contaminations in the binary response variable and significantly decrease the bias of the estimates by the MCGEE approach when the time period for cocaine to be cleared from urine is relatively short.

However, the difference between the bias of the MCGEE estimators and the GEE estimators shrinks significantly when the time period for cocaine to be cleared from urine increases from 1 day to 4 days. In other words, when h increases, the performances of the MCGEE model decreases, and so does the coverage probability of 95% CI. Since it is more difficult to accurately estimate the contamination indicators of each subject at each time point when $h = 4$, p_2 might be underestimated. As a result, p_1 is subjected to underestimation as well if we fail to detect contaminations for some subjects. In these scenarios, the marginal mean of the estimating equations may not be correctly specified, and the bias of the parameters' estimation of the MCGEE approach increases significantly.

After applying this approach to the SCU data, we observe that the treatment effect has been significantly changed when we adjust the marginal mean of the model by considering the contamination probability. It is anticipated that the bias of the treatment effect have been reduced from the original method. The study also reveals the feasibility of reducing the potential bias of the estimators through the detection of the report bias in the self-reported data and subsequent correction of the marginal

mean in the GEE model.

In conclusion, accurately modeling the repeated longitudinal binary outcomes with report bias depends on several factors. The key factor is to correctly estimate the contamination. The results of our study suggests that the proposed MCGEE approach performs well when the time period for cocaine to be cleared from urine is relatively short, e.g., 1 day. After increasing this time period to 4 days, the bias of the estimators of this approach increases, yet still outperforms the traditional GEE approach.

2.8 Appendix

From section 2.3, $U_N(\beta) = \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} (Z_i - \mu_i^*)$, define $\psi(Z_i; \beta) = D_i^{*\prime} V_i^{*-1} (Z_i - \mu_i^*)$. The Assumptions we need to prove the consistency and asymptotic normality of $\hat{\beta}_N$ are:

Assumption A. The subjects are independently sampled and there exists an upper bound $M < \infty$ such that the number of replicates $m_i < M$ for all subjects $i = 1, 2, ...$

Assumption B. There exists an upper bound $b < \infty$ such that $|X_i| < b$ for all subjects $i = 1, 2, ...$

Assumption C. It is assumed that $\frac{1}{N} \sum_{i=1}^{N} X_i X_i' \to B$ as $N \to \infty$, where B is a positive

definite matrix.

Define the matrices

$$I_0^*(\beta) = \lim_{N\to\infty} \frac{1}{N} \frac{\partial}{\partial \beta'} U_N(\beta) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (D_i^{*\prime} V_i^{*-1} D_i^*),$$

and

$$I_1^*(\beta) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (D_i^{*\prime} V_i^{*-1} A_i^* V_i^{*-1} D_i^*).$$

The existence of these limits is ensured by Assumption B. Moreover, Assumption C ensures that $I_0^*(\beta)$ and $I_1^*(\beta)$ are positive definite.

In order to prove the solution $\hat{\beta}_N$ of $U_N(\beta) = 0$ is consistent and asymptotic normally distributed for large N, we need to show that:

1. $\frac{1}{N} U_N(\beta_0) \to 0$ a.s., as $N \to \infty$.

2. $\frac{1}{N} \frac{\partial}{\partial \beta^T} U_N(\beta)$ converge uniformly to a nonstochastic limit which is nonsingular at β_0.

3. With probability one, $\psi(Z_i; \beta)$ are twice continuously differentiable with respect to $\beta \in B$, and $|\frac{\partial^2}{\partial \beta_j \partial \beta_k} \psi(Z_i; \beta)| < \infty$.

4. $|\psi(Z_i; \beta)| < \infty$, and $\frac{1}{\sqrt{N}} U_N(\beta_0) \underrightarrow{L} N(0, I_1^*(\beta_0))$.

Since our proof based on some assumptions and theorems from Yuan and Jennrich (1998), we verify their assumptions in our case in section 2.8.1, we prove the consistency of MCGEE estimator in section 2.8.2, and we show the asymptotic normality of MCGEE estimator in section 2.8.3.

2.8.1 Verifying the conditions

Yuan and Jennrich (1998) proved consistency and asymptotic normality of M estimators based on the following conditions, which are:

1. $\frac{1}{N}U_N(\beta_0) \to 0$ a.s., as $N \to \infty$.

2. There exists a neighborhood M of β_0 on which with probability one, all $\frac{1}{N}U_N(\beta)$ are continuously differentiable and $\frac{1}{N}\frac{\partial}{\partial \beta^T}U_N(\beta)$ converge uniformly to a nonstochastic limit which is nonsingular at β_0.

3.

$$\frac{1}{\sqrt{N}}U_N(\beta_0) \xrightarrow{L} N(0, I_1^*(\beta_0)),$$

as $N \to \infty$.

To prove that the MCGEE estimator is consistent and asymptotically normally distributed, we will demonstrate the conditions above are satisfied.

By Theorem 5 of Yuan and Jennrich (1998), to verify condition 1 of Juan and Jennrich

(1998) it suffices to verify the condition 4 of Yuan and Jennrich (1998):

4. For each i, $\psi(Z_i;\beta_0)$ has mean zero and variance-covariance matrix K_i, such that

$$\frac{1}{N}\sum_{i=1}^{N} K_i \to K,$$

for some positive-definite matrix K.

For self-reported data Z_i, since $E(Z_i) = \mu_i^*$, then $E(\psi(Z_i;\beta_0)) = 0$, and

$$var(\psi(Z_i;\beta_0)) = D_i^{*\prime} V_i^{*-1} A_i^* V_i^{*-1} D_i^*,$$

$$D_i^* = (1-2p_i) \otimes A_i X_i,$$

$$A_i^* = diag(var(Z_{i1}), ..., var(Z_{iT})),$$

$$var(Z_{it}) = (1-2p_{it})^2 \frac{e^{\beta' X_{it}}}{(1+e^{\beta' X_{it}})^2} + p_{it}(1-p_{it}),$$

then

$$A_i^* = diag((1-2p_{i1})^2 \frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} + p_{i1}(1-p_{i1}), ..., (1-2p_{iT})^2 \frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} + p_{iT}(1-p_{iT}))$$

Since $|X_i| < b < \infty$ for all $i = 1, 2, ...$by assumption, and the contamination proba-

bility, $0 \leq p_{it} \leq 1$, then

$$0 \leq (1 - 2p_i)^2 \leq 1,$$

$$0 \leq p_{it}(1 - p_{it}) \leq \frac{1}{4},$$

and

$$0 \leq \frac{e^{\beta' X_{it}}}{(1 + e^{\beta' X_{it}})^2} \leq \frac{1}{4}.$$

Therefore,

$$D_i^* = (1 - 2p_i) \otimes A_i X_i < \infty,$$

$$0 \leq var(Z_{it}) = (1 - 2p_{it})^2 \frac{e^{\beta' X_{it}}}{(1 + e^{\beta' X_{it}})^2} + p_{it}(1 - p_{it}) \leq \frac{1}{2},$$

and

$$V_i = A_i^{*\frac{1}{2}} C_i^*(\gamma) A_i^{*\frac{1}{2}} < \infty.$$

Hence, the variance-covariance matrix K_i of $\psi(\beta_0)$ satifies

$$\frac{1}{N} \sum_{i=1}^{N} K_i \to K,$$

for some positive-definite matrix K. Condition 1 of Juan and Jennrich (1998) has been verified.

To verify condition 2 of Juan and Jennrich (1998) it suffices to verify the following assumptions of Yuan and Jennrich (1998):

6. With probability one, $\psi(Z_i;\beta)$ are twice continuously differentiable with respect to $\beta \in B$.

7. For each $\beta \in B$,

$$\frac{1}{N}\sum_{i=1}^{N} E(\frac{\partial}{\partial \beta^T}\psi(Z_i;\beta)) \to I_0^*(\beta),$$

where $I_0^*(\beta)$ is nonsingular and with probability one

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \beta_t}\psi(Z_i;\beta) \to I_0^*(\beta),$$

as $N \to \infty$.

8. For each i,

$$|\frac{\partial^2}{\partial \beta_j \partial \beta_k}\psi(Z_i;\beta)| \leq S,$$

for some upper bound $S < \infty$.

Yuan and Jennrich proved that under conditions 6, 7, and 8, condition 2 is satisfied.

To verify condition 6, we have

$$\begin{aligned}\frac{\partial}{\partial \beta}\psi(Z_i;\beta) &= \frac{\partial}{\partial \beta}(D_i^{*\prime}V_i^{*-1}(Z_i - \mu_i^*)) \\ &= D_i^{*\prime}V_i^{*-1}D_i^* + (\frac{\partial}{\partial \beta}D_i^*)'V_i^{*-1}(Z_i - \mu_i^*) + D_i^{*\prime}(\frac{\partial}{\partial \beta}V_i^{*-1})(Z_i - \mu_i^*).\end{aligned}$$

Since $E(Z_i) = \mu_i^*$, the last two terms in the expression above have expectation zero,

so

$$E(\frac{\partial}{\partial\beta}\psi(Z_i;\beta)) = D_i^{*\prime}V_i^{*-1}D_i^*.$$

Moreover,

$$\frac{\partial}{\partial\beta}D_i^* = (1-2p_i)\otimes(\frac{\partial}{\partial\beta}A_i)X_i,$$

where $A_i = diag(var(Y_{i1}),...,var(Y_{iT}))$, and

$$\begin{aligned}\frac{\partial}{\partial\beta}\left(var(Y_{it})\right) =& \frac{\partial}{\partial\beta}\frac{e^{\beta' X_{it}}}{(1+e^{\beta' X_{it}})^2} \\ &= \frac{X_{it}e^{\beta' X_{it}}(1-e^{\beta' X_{it}})}{(1+e^{\beta' X_{it}})^3},\end{aligned}$$

$$\frac{\partial}{\partial\beta}A_i = diag\left(\frac{X_{i1}e^{\beta' X_{i1}}(1-e^{\beta' X_{i1}})}{(1+e^{\beta' X_{i1}})^3},...,\frac{X_{iT}e^{\beta' X_{iT}}(1-e^{\beta' X_{iT}})}{(1+e^{\beta' X_{iT}})^3}\right).$$

Since $|X_i| < b < \infty$, $0 \leq (1-2p_i) \leq 1$, $0 \leq \frac{e^{\beta' X_{it}}}{(1+e^{\beta' X_{it}})^2} \leq \frac{1}{4}$, and $0 \leq \frac{1}{1+e^{\beta' X_{it}}} \leq 1$,

$$\frac{\partial}{\partial\beta}D_i^* < \infty.$$

And,

$$\frac{\partial}{\partial\beta}V_i^{*-1} = -V_i^{*-1}(\frac{\partial}{\partial\beta}V_i^*)V_i^{*-1},$$

$$\frac{\partial}{\partial\beta}V_i^* = (\frac{\partial}{\partial\beta}A_i^{*\frac{1}{2}})C_i(\gamma)A_i^{*\frac{1}{2}} + A_i^{*\frac{1}{2}}C_i(\gamma)(\frac{\partial}{\partial\beta}A_i^{*\frac{1}{2}}).$$

$A_i^* = diag(var(Z_{i1}), ..., var(Z_{iT}))$, and

$$\begin{aligned}\frac{\partial}{\partial\beta}\left(var(Z_{it})\right) =& \frac{\partial}{\partial\beta}\left((1-2p_{it})^2\frac{e^{X_{it}\beta}}{(1+e^{X_{it}\beta})^2} + p_{it}(1-p_{it})\right)\\ =& (1-2p_{it})^2\frac{X_{it}e^{X_{it}\beta}(1-e^{X_{it}\beta})}{(1+e^{X_{it}\beta})^3}.\end{aligned}$$

Then

$$\begin{aligned}\frac{\partial}{\partial\beta}A_i^{*\frac{1}{2}} =& diag[\frac{\partial}{\partial\beta}\sqrt{(1-2p_{i1})^2\frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} + p_{i1}(1-p_{i1})}, ...,\\ & \frac{\partial}{\partial\beta}\sqrt{(1-2p_{iT})^2\frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} + p_{iT}(1-p_{iT})}]\\ =& \frac{1}{2}diag[\frac{(1-2p_{i1})^2\frac{X_{i1}e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} - 2\frac{X_{i1}e^{2\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^3}}{\sqrt{(1-2p_{i1})^2\frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} + p_{i1}(1-p_{i1})}}, ...,\\ & \frac{(1-2p_{iT})^2\frac{X_{iT}e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} - 2\frac{X_{iT}e^{2\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^3}}{\sqrt{(1-2p_{iT})^2\frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} + p_{iT}(1-p_{iT})}}]\\ =& \frac{1}{2}diag[\frac{(1-2p_{i1})^2\mu_{i1}(1-\mu_{i1})(1-2\mu_{i1})X_{i1}}{\sqrt{(1-2p_{i1})^2\frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} + p_{i1}(1-p_{i1})}}, ...,\\ & \frac{(1-2p_{iT})^2\mu_{iT}(1-\mu_{iT})(1-2\mu_{iT})X_{iT}}{\sqrt{(1-2p_{iT})^2\frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} + p_{iT}(1-p_{iT})}}].\end{aligned}$$

Therefore, $\frac{\partial}{\partial\beta}A_i^{*\frac{1}{2}} < \infty$, $\frac{\partial}{\partial\beta}V_i^* < \infty$, $(\frac{\partial}{\partial\beta}D_i^*)'V_i^{*-1}(Z_i-\mu_i^*) + D_i^{*\prime}(\frac{\partial}{\partial\beta}V_i^{*-1})(Z_i-\mu_i^*) < \infty$, and

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\beta}\psi(Z_i;\beta)=\frac{1}{N}\sum_{i=1}^{N}D_i^*V_i^{*-1}D_i^*.$$

Taking the second derivative,

$$\frac{\partial}{\partial\beta}(D_i^*V_i^{*-1}D_i^*)=(\frac{\partial}{\partial\beta}D_i^{*\prime})V_i^{*-1}D_i^*+D_i^{*\prime}(\frac{\partial}{\partial\beta}V_i^{*-1})D_i^*+D_i^{*\prime}V_i^{*-1}(\frac{\partial}{\partial\beta}D_i^*).$$

We also have

$$\frac{\partial}{\partial\beta}D_i^*=(1-2p_i)\otimes(\frac{\partial}{\partial\beta}A_i)X_i,$$

$$\frac{\partial}{\partial\beta}A_i=diag\left(\frac{X_{i1}e^{\beta'X_{i1}}(1-e^{\beta'X_{i1}})}{(1+e^{\beta'X_{i1}})^3},...,\frac{X_{iT}e^{\beta'X_{iT}}(1-e^{\beta'X_{iT}})}{(1+e^{\beta'X_{iT}})^3}\right).$$

Since $|X_i|<b<\infty$, $0\leq(1-2p_i)\leq 1$, $0\leq\frac{e^{\beta'X_{it}}}{(1+e^{\beta'X_{it}})^2}\leq\frac{1}{4}$, and $0\leq\frac{1}{1+e^{\beta'X_{it}}}\leq 1$,

$$\frac{\partial}{\partial\beta}D_i^*<\infty.$$

And

$$\frac{\partial}{\partial\beta}V_i^{*-1}=-V_i^{*-1}(\frac{\partial}{\partial\beta}V_i^*)V_i^{*-1},$$

$$\frac{\partial}{\partial\beta}V_i^* = (\frac{\partial}{\partial\beta}A_i^{*\frac{1}{2}})C_i(\gamma)A_i^{*\frac{1}{2}} + A_i^{*\frac{1}{2}}C_i(\gamma)(\frac{\partial}{\partial\beta}A_i^{*\frac{1}{2}}).$$

Then

$$\begin{aligned}\frac{\partial}{\partial\beta}A_i^{*\frac{1}{2}} &= diag[\frac{\partial}{\partial\beta}\sqrt{(1-2p_{i1})^2\frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} + p_{i1}(1-p_{i1})}, ..., \\ &\quad \frac{\partial}{\partial\beta}\sqrt{(1-2p_{iT})^2\frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} + p_{iT}(1-p_{iT})}] \\ &= \frac{1}{2}diag[\frac{(1-2p_{i1})^2\frac{X_{i1}e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} - 2\frac{X_{i1}e^{2\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^3}}{\sqrt{(1-2p_{i1})^2\frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} + p_{i1}(1-p_{i1})}}, ..., \\ &\quad \frac{(1-2p_{iT})^2\frac{X_{iT}e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} - 2\frac{X_{iT}e^{2\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^3}}{\sqrt{(1-2p_{iT})^2\frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} + p_{iT}(1-p_{iT})}}] \\ &= \frac{1}{2}diag[\frac{(1-2p_{i1})^2\mu_{i1}(1-\mu_{i1})(1-2\mu_{i1})X_{i1}}{\sqrt{(1-2p_{i1})^2\frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} + p_{i1}(1-p_{i1})}}, ..., \\ &\quad \frac{(1-2p_{iT})^2\mu_{iT}(1-\mu_{iT})(1-2\mu_{iT})X_{iT}}{\sqrt{(1-2p_{iT})^2\frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} + p_{iT}(1-p_{iT})}}].\end{aligned}$$

Therefore, $\frac{\partial}{\partial\beta}A_i^{*\frac{1}{2}} < \infty$, and $\frac{\partial}{\partial\beta}V_i^* < \infty$. Condition 6 is verified.

To verify condition 7 of Juan and Jennrich (1998) , the derivative of $\psi(Z_i;\beta)$ with respect to β is:

$$\begin{aligned}\frac{\partial}{\partial\beta}\psi(Z_i;\beta) =& \frac{\partial}{\partial\beta_t}(D_i^{*\prime}V_i^{*-1}(Z_i-\mu_i^*)) \\ &= D_i^*V_i^{*-1}D_i^* + (\frac{\partial}{\partial\beta}D_i^*)'V_i^{*-1}(Z_i-\mu_i^*) + D_i^{*\prime}(\frac{\partial}{\partial\beta}V_i^{*-1})(Z_i-\mu_i^*).\end{aligned}$$

We already showed the following equations when we verifying condition 6,

$$E(\frac{\partial}{\partial\beta}\psi(Z_i;\beta)) = D_i^* V_i^{*-1} D_i^*.$$

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\beta}\psi(Z_i;\beta) = \frac{1}{N}\sum_{i=1}^{N} D_i^* V_i^{*-1} D_i^*.$$

To complete verifying condition 7, we need to show that

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\beta}\psi(Z_i;\beta) \to I_0^*(\beta).$$

almost surely as $N \to \infty$.

Since

$$\frac{1}{N}\sum_{i=1}^{N} D_i^* V_i^{*-1} D_i^* = \frac{1}{N}\sum_{i=1}^{N}((1-2p_i)\otimes A_i X_i')' V_i^{*-1}(1-2p_i)\otimes A_i X_i',$$

and $(1-2p_i)$, V_i, A_i are all bounded from previous proof. Then $((1-2p_i)\otimes A_i)' V_i^{*-1}(1-2p_i)\otimes A_i$ is bounded below by a positive constant b_i.

Let a denote any $T \times 1$ vector, then

$$\frac{1}{N}a'\sum_{i=1}^{N} X_i((1-2p_i)\otimes A_i)' V_i^{*-1}(1-2p_i)\otimes A_i X_i' a \geq \frac{1}{N} b_i a' \sum_{i=1}^{N} X_i X_i' > 0,$$

by Assumption C, which is

$$\frac{1}{N}\sum_{i=1}^{N} X_i X_i' \to B,$$

as $N \to \infty$, where B is a positive definite matrix.

Then,

$$\frac{\partial}{\partial\beta}\psi(Z_i;\beta) < \infty,$$

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial\beta}\psi(Z_i;\beta) \to I_0^*(\beta),$$

almost surely as $N \to \infty$.

To verify condition 8 of Juan and Jennrich (1998), we already show that each term of the second derivatives of $\psi(Z_i;\beta)$ with respect to β is bounded when we verify condition 6 ($\frac{\partial}{\partial\beta}D_i^* < \infty$, $\frac{\partial}{\partial\beta}V_i^{*-1} < \infty$).

Hence,

$$\frac{\partial^2}{\partial\beta\partial\beta}\psi(Z_i;\beta) < \infty.$$

In conclusion, condition 2 of Juan and Jennrich (1998) has been verified.

Liapounov's Theorem and Cramer-Wald Theorem are used to verify condition 3 of

Juan and Jennrich (1998),

$$\frac{1}{\sqrt{N}} U_N(\beta_0) \xrightarrow{L} N(0, I_1^*(\beta_0)),$$

as $N \to \infty$.

As defined earlier,

$$U_N(\beta_0) = \sum_{i=1}^{N} \psi(Z_i; \beta_0) = \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} (Z_i - \mu_i^*).$$

Let a denote any $T \times 1$ vector, to apply Liapounov's Theorem, take

$$r_i = a' D_i^{*\prime} V_i^{*-1} Z_i.$$

Then the mean of r_i is

$$m_i = E(r_i) = a' D_i^{*\prime} V_i^{*-1} \mu_i^*,$$

and the variance of r_i is

$$Var(r_i) = a' D_i^{*\prime} V_i^{*-1} A_i^* V_i^{*-1} D_i^* a.$$

Define

$$c_n^2 = \sum_{i=1}^{N} Var(r_i) = \sum_{i=1}^{N} a' D_i^{*\prime} V_i^{*-1} A_i^* V_i^{*-1} D_i^* a = O(N),$$

since

$$\frac{1}{N}\sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} A_i^* V_i^{*-1} D_i^* \to I_1^*,$$

under condition 4.

Assume $E(|Z_i - \mu_i|^3) = \mu_{3i}^* < \infty$. Taking $\delta = 1$, the third central moment is:

$$\begin{aligned} E(|r_i - m_i|^3) &= E(|a' D_i^{*\prime} V_i^{*-1}(Z_i - \mu_i^*)|^3) \\ &\leq (a' D_i^{*\prime} V_i^{*-1})^3 E(|Z_i - \mu_i^*|^3) \\ &= (a' D_i^{*\prime} V_i^{*-1})^3 \mu_{3i}^*. \end{aligned}$$

So

$$\sum_{i=1}^{N} E(|r_i - m_i|^3) = O(N),$$

since D_i^* and V_i^* are bounded, which have been showed when verifying condition 4 of Yuan and Jennrich(1998).

Then

$$\frac{1}{c_n^3}\sum_{i=1}^{N} E(|r_i - m_i|^3) = \frac{O(N)}{O(N^{\frac{3}{2}})} = O(N^{-1/2}),$$

which converges to zero as $N \to \infty$. Therefore, the conditions of Liapounov's theo-

rem are satisfied, and

$$\begin{aligned} T_N &= \frac{\sum_{i=1}^{N}(r_i - m_i)}{c_n} \\ &= \frac{\sum_{i=1}^{N} a' D_i^{*\prime} V_i^{*-1}(Z_i - \mu_i^*)}{\sqrt{\sum_{i=1}^{N} a' D_i^{*\prime} V_i^{*-1} A_i^* V_i^{*-1} D_i^* a}} \\ &\xrightarrow{L} N(0,1), \end{aligned}$$

as $N \to \infty$.

By Slutsky's Theorem,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} a' D_i^{*\prime} V_i^{*-1}(Z_i - \mu_i^*) \xrightarrow{L} N(0, a' I_1^*(\beta) a).$$

By the Cramer-Wold Theorem,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1}(Z_i - \mu_i^*) \xrightarrow{L} N(0, I_1^*(\beta)),$$

where $I_1^*(\beta) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (D_i^{*\prime} V_i^{*-1} A_i^* V_i^{*-1} D_i^*)$.

Thus, condition 3 has been verified.

2.8.2 Proof of consistency

Theorem 1. *Under Assumptions A-C, with probability one there exist zeros* $\hat{\beta}_N$ *of* $U_N(\beta) = 0$ *such that* $\hat{\beta}_N \to \beta_0$ *as* $N \to \infty$.

By condition 1 of Yuan and Jennrich (1998), $\frac{1}{N} U_N(\beta_0) \to 0$ a.s., as $N \to \infty$. And by condition 2 of Yuan and Jennrich (1998), $\frac{1}{N} \frac{\partial}{\partial \beta'} U_N(\beta)$ is nonsingular at β_0. Thus, β_0 is the unique zero of $U(\beta)$ in a neighborhood M of β_0.

Theorem 1 of Yuan and Jennrich (1998) states that under conditions 1 and 2, for any $\delta > 0$, there exists $\hat{\beta}_N \in M(\beta_0, \delta)$ such that $U_N\left(\hat{\beta}_N\right) = 0$ with probability 1, for all N sufficiently large.

By Theorem 1, there exists a zero $\hat{\beta}_N$ of $U_N(\beta)$ in $M(\beta_0, \delta)$ for all N sufficiently large. Let β^* be any limit point of $\hat{\beta}_N$, then $\beta^* \in M(\beta_0, \delta)$. Let $\hat{\beta}_{N_i}$ be any subsequence of $\hat{\beta}_N$, then $\hat{\beta}_{N_i} \to \beta^*$. Thus, $U_{N_i}\left(\hat{\beta}_{N_i}\right) \to U(\beta^*)$, and $U(\beta^*) = 0$.

Since β_0 is the only zero of $U(\beta)$ in a neighborhood $M(\beta_0, \delta)$, then $\beta^* = \beta_0$. Since this is true for all limit points of $\hat{\beta}_N$, $\hat{\beta}_N \to \beta_0$. Since conditions 1 and 2 hold with probability one, $\hat{\beta}_N \to \beta_0$ with probability one.

2.8.3 Proof of asymptotic normality

Theorem 2. *Under Assummptions A-C,* $\sqrt{N}\left(\hat{\beta}-\beta_0\right) \underrightarrow{L} N\left(0, I_0^{*-1}(\beta_0) I_1^*(\beta_0) I_0^{*-1}(\beta_0)\right)$, *as* $N \to \infty$.

To prove **Theorem 2**, a Taylor series expansion of $\frac{1}{N} U_N\left(\hat{\beta}_N\right) = \frac{1}{N} \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1}\left(Z_i - \mu_i^*\right)$ at β_0 yields

$$U_N\left(\hat{\beta}_N\right) = U_N\left(\beta_0\right) + \frac{\partial}{\partial \beta'} U_N\left(\beta_0\right)\left(\hat{\beta}_N - \beta_0\right) = 0.$$

Setting the expression above, and rearranging the terms, we get

$$\sqrt{N}\left(\hat{\beta}_N - \beta_0\right) \cong -\left(\frac{1}{N}\frac{\partial}{\partial \beta'} U_N\left(\beta_0\right)\right)^{-1} \frac{1}{\sqrt{N}} U_N\left(\beta_0\right).$$

To prove **Theorem 1**, we have already demonstrated that:

$$\frac{1}{N}\frac{\partial}{\partial \beta'} U_N\left(\beta_0\right) \to I_0^*(\beta_0),$$

almost surely as $N \to \infty$.

Since we already show that when verifying condition 3 of Yuan and Jennrich (1998),

$$\frac{1}{\sqrt{N}} U_N\left(\beta_0\right) \underrightarrow{L} N\left(0, I_1^*(\beta_0)\right).$$

as $N \to \infty$.

Then by Theorem 4 of Yuan and Jennrich (1998) and by slutsky's Theorem,

$$\sqrt{N}\left(\hat{\beta}-\beta_0\right) \underrightarrow{L} N\left(0, I_0^{*-1}(\beta_0)I_1^*(\beta_0)I_0^{*-1}(\beta_0)\right),$$

as $N \to \infty$.

And $\hat{I}_0^*$ and $\hat{I}_1^*$ can be estimated as:

$$\begin{aligned}\hat{I}_0^* &= \sum_{i=1}^{N}(\hat{D}_i^{*\prime}\hat{V}_i^{*-1}\hat{D}_i^*) \\ &= \sum_{i=1}^{N}(1-2p_i)\otimes X_i'\hat{A}_i(\hat{A}_i^{*\frac{1}{2}}\hat{C}_i^*(\gamma)\hat{A}_i^{*\frac{1}{2}})^{-1}(1-2p_i)\otimes\hat{A}_iX_i,\end{aligned}$$

$$\begin{aligned}\hat{I}_1^* &= \sum_{i=1}^{N}(\hat{D}_i^{*\prime}\hat{V}_i^{*-1}\hat{A}_i^*\hat{V}_i^{*-1}\hat{D}_i^*) \\ &= \sum_{i=1}^{N}(1-2p_i)\otimes X_i'\hat{A}_i(\hat{A}_i^{*\frac{1}{2}}\hat{C}_i^*(\gamma)\hat{A}_i^{*\frac{1}{2}})^{-1}\hat{A}_i^*(\hat{A}_i^{*\frac{1}{2}}\hat{C}_i^*(\gamma)\hat{A}_i^{*\frac{1}{2}})^{-1}(1-2p_i)\otimes\hat{A}_iX_i.\end{aligned}$$

2.9 References

Argoff C.E., McCleane G. Pain Management Secrets. Mosby/Elsevier. 2009.

Babor T.F., Del Boca F.K., McRee B. Estimating measurement error in alcohol dependence symptomatology: findings from a multisite study. Drug Alcohol Depend.

1997; 45(1-2):13-20.

Babor T.F., Steinberg K., Anton R., Del Boca F. Talk is cheap: Measuring drinking outcomes in clinical trials. J Stud Alcohol. 2000; 61(1):55-63.

Beck A. T., Wright F., Newman C., Liese B. Cognitive therapy of substance abuse. New York, NY: Guilford Press.1993.

Biemer P. , Trewin D. A review of measurement error effects on the analysis of survey data. In Survey Measurement and Process Quality. New York: Wiley. 1997.

Blattmana C., Jamisonb J., Koroknay-Paliczc T., Rodriguesd K., Sheridane M. Measuring the measurement error: A method to qualitatively validate survey data. Journal of Development Economics. 2016;120:99-112.

Brown J., Kranzler H.R., Del Boca F.K. Self-reports by alcohol and drug abuse inpatients: Factors affecting reliability and validity. British Journal of Addiction. 1992; 87(7):1013-1024.

Brown E.E., Finlay J.M., Wong J.T., Damsma G., Fibiger H.C. Behavioral and neurochemical interactions between cocaine and buprenorphine: implications for the pharmacotherapy of cocaine abuse. J Pharmacol Exp Ther. 1991; 256:119-125.

Carroll KM. A Cognitive Behavioral Approach: Treating Cocaine Addiction. Rockville, MD: National Institute on Drug Abuse; 1998.

Del Boca F.K., Darkes J. The validity of self-reports of alcohol consumption: State of the science and challenges for research. Addiction. 2003; 98 (2):1-12.

Del Boca F.K., Noll J.A. Truth or consequences: The validity of self-report data in health services research on addictions. Addiction.2000; 95(3):347-360.

Digiusto E., Seres V., Bibby A., Batey R. Concordance between urinalysis results and self-reported drug use by applicants for methadone maintenance in Australia. Addictive Behaviors.1996; 21(3):319-329.

Fiellin D.A., Barry D.T., Sullivan L.E., Cutter C.J., Moore B.A., O'Connor P.G., Schottenfeld R.S. A randomized trial of cognitive behavioral therapy in primary care-based buprenorphine. Am J Med. 2013;126(1):74.

Fitzmaurice G.M., Davidian M, Verbeke G, Molenberghs G. Longitudinal Data Analysis. Chapman Hall/CRC: Boca Raton, FL, 2009.

Gonzalez V. M., Schmitz J. M., DeLaune K. A. The role of homework in cognitive-behavioral therapy for cocaine dependence. Journal of Consulting and Clinical Psychology. 2006; 74:633-637.

Johnson T., Fendrich M.Modeling Sources of Self-report Bias in a Survey of Drug Use Epidemiology. Annals of Epidemiology. 2005; 15(5): 381-389.

Kamien J. H. B., Spealman R. D. Modulation of the discriminative-stimulus effects of cocaine by buprenorphine. Behav Pharmacol. 1991; 2: 517-520.

Liang K., Zeger S. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73:13-22.

Ling W., Hillhouse M.P., Saxon A.J., Mooney L.J., Thomas C.M., Ang A., Matthews A.G., Hasson A., Annon J., Sparenborg S., Liu D.S., McCormack J., Church S., Swafford W., Drexler K., Schuman C., Ross S., Wiest K., Korthuis P.T., Lawson W., Brigham G.S., Knox P.C., Dawes M., Rotrosen J. Buprenorphine + naloxone plus naltrexone for the treatment of cocaine dependence: the Cocaine Use Reduction with Buprenorphine (CURB) study. Addiction. 2016; 111(8):1416-27.

Lloyd-Smith E., Rachlis B.S., Tobin D., Stone D., Li K., Small W., Wood E., Kerr T. Assisted injection in outdoor venues: an observational study of risks and implications for service delivery and harm reduction programming. Harm Reduction Journal. 2010; 7: 6-10.

Macdonald S., Erickson P., Wells S., Hathaway A., Pakula B. Predicting violence

among cocaine, cannabis, and alcohol treatment clients. Addict. Behav. 2008; 33:201-205.

Magill M., Ray L.A. Cognitive-behavioral treatment with adult alcohol and illicit drug users: a meta-analysis of randomized controlled trials. J Stud Alcohol drugs. 2009;70:516-527.

Magura S., Goldsmith D., Casriel C., Goldstein P.J., Lipton D.S. The validity of methadone clients' self-reported drug use. International Journal of the Addictions. 1987; 22 (8): 727-749.

Maude-Griffin P.M., Hohenstein J.M., Humfleet G.L., Reilly P.M., Tusel D.J., Hall S.M. Superior efficacy of cognitive-behavioral therapy for urban crack cocaine abusers: main and matching effects. J. Consult Clin. Psychol. 1998; (66):832-837.

Mooney L.J., Nielsen S., Saxon A., Hillhouse M., Thomas C., Hasson A., Stablein D., McCormack J., Lindblad R., Ling W. Cocaine use reduction with buprenorphine (CURB): rationale, design, and methodology. Contemporary Clinical Trials. 2013; 34 (2): 196-204.

Myrick H., Henderson S., Dansky B., Pelic C., Brady K.T.Clinical characteristics of under-reporters on urine drug screens in a cocaine treatment study. The American Journal on Addictions. 2002; 11(4):255-261.

Qureshi A.I., Suri M.F., Guterman L.R., Hopkins L.N. Cocaine use and the likelihood of nonfatal myocardial infarction and stroke: data from the Third National Health and Nutrition Examination Survey. Circulation. 2001;103: 502-506.

Rawson R.A., McCann M.J., Flammino F., Shoptaw S., Miotto K., Reiber C., Ling W. A comparison of contingency management and cognitive-behavioral approaches for stimulant- dependent individuals. Addiction. 2006; 101(2): 267-274.

Sherman M.F., Bigelow G.E. Validity of patients' self-reported drug use as a function of treatment status. Drug and Alcohol Dependence. 1992; 30(1):1-11.

Sofuoglu M., Kosten T.R. Emerging pharmacological strategies in the fight against cocaine addiction. Expert Opin. Emerg. Drugs. 2006; (11): 91-98.

Solbergsdottir E., Bjornsson G., Gudmundsson L.S., Tyrfingsson T., Kristinsson J. Validity of self-reports and drug use among young people seeking treatment for substance abuse or dependence. J Addict Dis. 2004; 23(1):29-38.

Umbricht A., DeFulio A., Winstanley E.L., Tompkins D.A., Peirce J., Mintzer M.Z., Strain E.C., Bigelow G.E. Topiramate for cocaine dependence during methadone maintenance treatment: A randomized controlled trial. Drug Alcohol Depend. 2014;140:92-100.

Walsh S.L., Middleton L.S., Wong C.J., Nuzzo P.A., Campbell C.L., Rush C.R., Lofwall M.R. Atomoxetine does not alter cocaine use in cocaine dependent individuals: double blind randomized trial. Drug Alcohol Depend. 2013;130(1-3):150-7.

Wedderburn R.W.M. Quasi-likelihood functions, generalized liner models, and the Gauss-Newton method. Biometrika. 1974; 61:439-447.

Wilcox C.E., Bogenschutz M.P., Nakazawa M., Woody G. Concordance between self-report and urine drug screen data in adolescent opioid dependent clinical trial participants. Addictive Behaviors. 2013; 38(10):2568-74.

Winhusen T.M., Somoza E.C., Singal B., Kim S., Horn P.S., Rotrosen J. Measuring outcome in cocaine clinical trials: a comparison of sweat patches with urine toxicology and participant self-report. Addiction. 2003; 98(3):317-24.

Yuan K.H., Jennrich R.I. Asymptotics of Estimating Equations under Natural Conditions. Journal of Multivariate Analysis. 1998; 65: 245-260.

Chapter 3

Weighted Generalized Estimating Equations Approach for Longitudinal Binary Outcomes with Significant Report Bias

3.1 Introduction

In the previous Chapter, in order to correct the report bias in self-reported daily drug use, we used the urine test results to detect the contamination, and estimated the true marginal means of the self-reported binary outcomes. We proposed Mean Corrected Generalized Estimating Equations (MCGEE) to apply the contamination estimation to correct the marginal means in the GEE model of self-reported data.

The bias of the estimators have significantly dropped in comparison to the Generalized Estimating Equations (GEE). However, all these results rely on the validity of the estimation of the contamination probability. When the time period for cocaine to be cleared from urine increases, the accuracy of the estimation of contamination probability decreases, and hence, the bias of the MCGEE model's estimators increase. Therefore we propose the following: include a weight function of the contamination probability into the MCGEE approach and build a Mean Corrected Weighted Generalized Estimating Equations (MCWGEE) approach to further reduce the potential bias of the estimators.

The Weighted Generalized Estimating Equations (WGEE) is an extension of the GEE approach. WGEE is often used for analyzing incomplete longitudinal data, and gives consistent estimations under Missing at Random (MAR) when the dropout mechanism is correctly specified. It was first proposed by Robins et al. (1995), in the form of the inverse probability weighted generalized estimating equation. In the approach named "observation-specific weighted method", each observation is weighted by the inverse probability of being observed. In some scenarios, this approach can be more efficient than the "subject-specific weighted method" (Fitzmaurice et al. 1995), which assigns a single weight that applies to all the observations from each time point of the same subject (O'Hara et al. 1999).

Under the assumption of MAR, if the mean model and the missing mechanism of the model are correctly specified, the WGEE method provides a consistent estimate

of regression parameters. Robins et al. (1995) has proposed WGEE of the form:

$$\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}' V_{it}^{-1} W_{it} \left(Y_{it} - \mu_{it}\right) = 0,$$

where D_{it}, μ_{it}, and V_{it} have the same form as previously defined in Chapter 2. $W_{it} = diag\left(r_{it} w_{it}\right)$ is the weighted matrix, where $r_{it} = 1$ if the outcome for subject i is observed at time t; otherwise, $r_{it} = 0$. As a result, the weight W_{it} is w_{it} for an observed visit and 0 for an unobserved visit.

Preisser et al. compared the WGEE and GEE approaches for repeated binary outcomes with MAR using a simulation study, in which the WGEE provided a smaller finite sample bias than the GEE. Moreover, the WGEE with observation-specific weight provided more accurate estimates than the WGEE with subject-specific weight (Preisser et al. 2002). Lipsitz et al. conducted another simulation study for the analysis of a similar binary response outcome with missing data, and concluded that the GEE model performed well under Missing Completely at Random (MCAR), and underperformed under MAR. On the other hand, the bias of the WGEE approach is negligible under MAR (Lipsitz et al. 2000). Other studies compared the WGEE estimators and the weighted least squares estimators under the MAR assumption through simulation, and suggested that the WGEE outperformed the weighted least squares estimators, and remained consistent under various scenarios of missing data and sample size (Lin et al. 2006).

There are two types of weights usually used in the WGEE weight assignment. First, the observation-specific weight proposed by Preisser et al.(2002), can be obtained through a logistic regression model. The weight can later be used in the WGEE approach for parameter estimations. Under MAR missing mechanism, let R_{it} be the indicator of observing the outcome at time t, and $\lambda_{it} = P(R_{it} = 1|R_{i(t-1)} = 1, X_{it}, Y_{it}, \theta)$ be the probability of observing the outcome at time t for the ith individual conditional on the individual being observed at the previous time point $t-1$. For the first time point, assume $R_{i1} = 1$ and $\lambda_{i1} = 1$.

Intuitively, $\hat{\lambda}_{it}$ can be estimated by fitting a logistic model, $logit(\lambda_{it}(\theta)) = Z_{it}\theta$, with a vector of predictors, Z_{it}, which may include indicator variables of visit, covariates, and past response variables. w_{it} is then defined as the inverse of the unconditional probability of being observed at time t, which can be estimated by the conditional probability,

$$\hat{w}_{it} = \left(\hat{\lambda}_{i1} \times ... \times \hat{\lambda}_{it}\right)^{-1}.$$

In this approach, an observation with a low probability of being observed will receive a large weight.

Lipsitz et al. considered the aforementioned WGEE approach with the observation-specific weight for handling missing response data. They concluded that this approach yielded less bias than the standard GEE approach or the multiple imputation approach under MAR (Lipsitz et al. 2000). In another study that used the inverse probability of the observation level WGEE model for data with non-ignorable non-

responses, the simulation produced mostly unbiased estimates (Troxel et al. 1997). Paik and Wang proposed an alternative observation level weighting approach for a longitudinal study with data missing, where weight is a decreasing function of an artificially created observation indicator. Their work showed that the proposed method yielded predominately unbiased and efficient results (Paik and Wang 2009). Further, another study investigated a class of inverse intensity-of-visit process-weighted estimators in marginal regression models for longitudinal responses that was observed in continuous time, and showed that the consistency still holds (Lin et al. 2004).

The second one is the subject-specific weight, which is sometimes referred to as the cluster-level weight. It assigns a single weight to each subject, i.e. all the observations from the same subject receive the same weight. Under the MAR scenario, a subject's weight w_i is the inverse of the probability of dropping out at the observed time of dropout.

$$\hat{w}_i^{-1} = \left(\prod_{t=2}^{m-1} \hat{\lambda}_{it} \right) \left(1 - \hat{\lambda}_{im} \right)^{I(m \leq T)},$$

where λ_{it} is obtained following the approach in the previous paragraph, m is the time of drop-out, I is the indicator function (Fitzmaurice et al. 1995).

Huang and Leroux performed a study using the cluster-level WGEE with weights determined by the size of the cluster as well as the subject's characteristic within the cluster. Their study yielded satisfactory results (Huang and Leroux 2011).

O'Hara Hines et al. compared two different weighted approaches of the WGEE model for joint continuous and categorical responses from clustered data with MAR missing mechanisms. The subject-specific weight approach was less efficient and more sensitive to influential observations (O'Hara Hines et al 1999). Preisser et al. reported a similar finding, suggesting that the subject-specific weight approach was considered less efficient than the observation-specific approach especially with minimal dropouts (Preisser et al. 2002). One possible reason may be that some subjects with limited information at the first few time points might have received smaller weight than they should. Therefore, this approach yielded a larger bias when dropout was minimal.

Despite the fact that various studies have used the WGEE approach to model the incomplete longitudinal data with dropouts under MAR assumption, rarely have any applied this method to model data with contamination. One possible explanation is that missing data can be easily documented, while contamination is more difficult to detect. In the Self-reported Cocaine use and Urine test (SCU) data, urine test results can be used to detect the contamination in self-reported outcomes, providing guidance on assigning the weight subsequently used in the WGEE model.

As such, we explore an alternative approach, in which we borrow the framework from the WGEE methods to estimate the treatment effect among adults with cocaine dependence. In our approach, the "contamination probability in self-reported data" is not only used to adjust the marginal means of data with report bias, but also to estimate the weight in the Mean Corrected WGEE model, i.e., the weight

assignment and marginal means estimation are based on the information from self-reported and urine binary data. In this chapter, we consider both observation-specific weighted approach and subject-specific weighted approach, and compare their performances with MCGEE approach through several different settings of sample sizes, contamination probabilities, and the time period for cocaine to be cleared from urine.

This chapter is organized as follows. In section 3.2, we give the notation and model equations. We study the asymptotic properties of the estimators from our proposed MCWGEE approach in section 3.3. The bias of the estimators when data is contaminated is explored in section 3.4. The performance of the MCWGEE approach on finite sample data is evaluated through simulation studies in section 3.5. In section 3.6, we analyze the SCU data using the MCWGEE methods. Finally, we provide the discussion and conclusion in section 3.7.

3.2 Methods

3.2.1 Weighted generalized estimating equations

To correct bias resulting from the contamination in the self-reported measures, we borrow the idea of the inverse probability weight from the WGEE approach. From Chapter 2, Y_{it} is the true drug use variable, and X_{it} is the covariate vectors for estimation at times $t = 1, ..., T$ for subjects $i = 1, ..., N$. Then, for the ith subject at time t, $Y_{it} = 1$ if the subject uses drug, $Y_{it} = 0$ otherwise. Y_{it} is a binary response

variable and its marginal distribution is Bernoulli:

$$f_y\left(y_i \mid X_i\right) = pr\left(Y_{i1} = y_1, ..., Y_{iT} = y_T \mid X_i\right) = exp(y_{it}\eta_{it} - log(1 + exp(\eta_{it}))).$$

The marginal mean of the true drug use for the ith subject at a given time point t is denoted by μ_{it}. Let β be a vector of the regression parameters, then

$$\mu_{it} = E\left(Y_{it} \mid X_i, \beta\right) = Pr\left(Y_{it} = 1 \mid X_i, \beta\right),$$

and with a logit link we will have

$$\eta_{it} = \log \frac{\mu_{it}}{1 - \mu_{it}} = x_{it}\beta.$$

The GEE form of Y_{it} is:

$$U_\beta\left(\beta\right) = \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it}' V_{it}^{-1}\left(Y_{it} - \mu_{it}\right) = 0,$$

where $D_{it} = \partial\mu_{it}/\partial\beta$, and V_i is the covariance matrix of Y_i, which can be decomposed into the form $A_i^{\frac{1}{2}} C_i(\gamma) A_i^{\frac{1}{2}}$, where A_i is a matrix with the marginal variances on the main diagonal and zeros elsewhere, γ is a vector which fully characterizes $C_i(\gamma)$, which serves as a working correlation matrix of Y_i's.

After forming μ_{it} to a vector, $\mu_i = E(Y_i|X_i, \beta) = (\mu_{i1}, ...\mu_{iT})'$, and since we assumed

$\eta_i = \log \frac{\mu_i}{1-\mu_i} = x_i\beta$, then

$$D_i = \partial\mu_i/\partial\beta = \frac{e^{x_i\beta}}{(1+e^{x_i\beta})^2} X_i.$$

And $A_i = diag(var(Y_{i1}), ..., var(Y_{iT}))$, $var(Y_{it}) = \mu_{it} \times (1-\mu_{it}) = \frac{e^{x_{it}\beta}}{(1+e^{x_{it}\beta})^2}$, then D_i can be written as:

$$D_i = \partial\mu_i/\partial\beta = A_i X_i.$$

Thus, we can write the GEE of the true drug use of the form:

$$U_\beta(\beta) = \sum_{i=1}^{N} X_i' A_i (A_i^{\frac{1}{2}} C_i(\gamma) A_i^{\frac{1}{2}})^{-1} (Y_i - \mu_i) = 0.$$

Let the variable R_{it} represent outcome contamination at times $t = 1, ..., T$, for subjects $i = 1, ..., N$, suggesting whether self-reported data is the same as true drug use. $R_{it} = 1$ if there is contamination, i.e. self reported data is not the same as true drug use data, otherwise $R_{it} = 0$. Let Z_{it} denote the self reported drug use, then

$$Z_{it} = Y_{it}(1 - R_{it}) + (1 - Y_{it}) R_{it}.$$

The estimating equation derived from the inverse probability WGEE approach for

self-reported data is:

$$U_\beta(\beta) = \sum_{i=1}^{N} D_i' V_i^{-1} W_i (Z_i - \mu_i).$$

Compared with the equation from the GEE approach, Y_i has been replaced by Z_i, $D_i = \partial\mu_i/\partial\beta$, V_i is the covariance matrix of Y_i, which equals $A_i^{\frac{1}{2}} C_i(\gamma) A_i^{\frac{1}{2}}$, and $W_i = diag(\frac{1}{w_{i1}}, ..., \frac{1}{w_{iT}})$, where $\frac{1}{w_{it}}$ is the weight for the ith individual at the tth time point. The weight is defined as the inverse of the contamination probability.

However, since the mean of self-reported data is not equal to the true marginal mean, i.e.,

$$E(Z_i) \neq \mu_i,$$

the previous WGEE equation may not be equal to 0, and may not give unbiased estimators. Therefore, we follow the approach discussed in Chapter 2, and propose a Mean Corrected WGEE (MCWGEE) approach in section 3.2.2.

3.2.2 Mean corrected weighted generalized estimating equations

Following the notation in Chapter 2, the expected value of Z_{it} is denoted by μ_{it}^*, and the expected value of the variable for contamination R_{it} is denoted by p_{it}. And we assume that the true drug use variable Y_{it} and the variable for contamination R_{it} are independent given X_{it}. Then, μ_{it}^* can be calculated as:

$$\begin{aligned}\mu_{it}^{*} = E(Z_{it}|X_{it},\beta) &= E(Y_{it}|X_{it},\beta) \times E((1-R_{it})\,|X_{it},\beta) + E((1-Y_{it})\,|X_{it},\beta) \times E(R_{it}|X_{it},\beta) \\ &= \mu_{it} \times E((1-R_{it})\,|X_{it},\beta) + (1-\mu_{it}) \times E(R_{it}|X_{it},\beta) \\ &= \mu_{it} - 2\mu_{it} \times p_{it} + p_{it}.\end{aligned}$$

After forming μ_{it}^{*} to a vector, $\mu_{i}^{*} = (\mu_{i1}^{*}, ...\mu_{iT}^{*})'$,

$$\mu_{i}^{*} = E(Z_{i}|X_{i},\beta),$$

the MCWGEE form of the self-reported data Z_i can be written as:

$$U_{\beta}^{*}(\beta) = \sum_{i=1}^{N} D_{i}^{*\prime} V_{i}^{*-1} W_{i}\,(Z_{i} - \mu_{i}^{*}) = 0.$$

And,

$$\begin{aligned}D_{i}^{*} = \partial\mu_{i}^{*}/\partial\beta &= (1-2p_{i}) \otimes \frac{\partial\mu_{i}}{\partial\beta} \\ &= (1-2p_{i}) \otimes \frac{e^{x_{i}\beta}}{(1+e^{x_{i}\beta})^{2}} X_{i} \\ &= (1-2p_{i}) \otimes A_{i}X_{i},\end{aligned}$$

where $A_i = diag(var(Y_{i1}), ..., var(Y_{iT}))$, $var(Y_{it}) = \mu_{it} \times (1-\mu_{it}) = \frac{e^{x_{it}\beta}}{(1+e^{x_{it}\beta})^{2}}$. $\otimes$ means only multiply the tth row of vector $1-2p_i$ by the same tth row of matrix A_iX_i, i.e., $(1-2p_i) \otimes A_iX_i = ((1-2p_{i1}) \times a_{i1}x_{i1}, ..., (1-2p_{iT}) \times a_{iT}x_{iT})'$.

$W_i = diag(\frac{1}{w_{i1}}, ..., \frac{1}{w_{iT}})$, where $\frac{1}{w_{it}}$ is the weight for the ith individual at the tth time point. V_i^* is the covariance matrix of Z_i, which can be decomposed into the form $A_i^{*\frac{1}{2}} C_i^*(\gamma) A_i^{*\frac{1}{2}}$, in which A_i^* is a matrix with marginal variances on the main diagonal and zeros elsewhere, i.e., $A_i^* = diag(var(Z_{i1}), ..., var(Z_{iT}))$, and

$$\begin{aligned} var(Z_{it}) &= \mu_{it}^*(1-\mu_{it}^*) \\ &= (\mu_{it} - 2\mu_{it}p_{it} + p_{it})(1 - \mu_{it} + 2\mu_{it}p_{it} - p_{it}) \\ &= \mu_{it} - 2\mu_{it}p_{it} + p_{it} - \mu_{it}^2 + 2\mu_{it}^2 p_{it} - p_{it}\mu_{it} + 2\mu_{it}^2 p_{it} - 4\mu_{it}^2 p_{it}^2 + 2\mu_{it}p_{it}^2 - \mu_{it}p_{it} \\ &\quad + 2\mu_{it}p_{it}^2 - p_{it}^2 \\ &= (1-2p_{it})^2 \mu_{it}(1-\mu_{it}) + p_{it}(1-p_{it}) \\ &= (1-2p_{it})^2 var(Y_{it}) + p_{it}(1-p_{it}). \end{aligned}$$

γ is a vector which fully characterize $C_i^*(\gamma)$, which is a working correlation matrix of Z_i's.

Hence, we can write the self-reported data's MCWGEE of the form:

$$U_\beta^*(\beta) = \sum_{i=1}^{N} (1-2p_i) \otimes X_i' A_i (A_i^{*\frac{1}{2}} C_i^*(\gamma) A_i^{*\frac{1}{2}})^{-1} W_i \left(Z_i - (\mu_i - 2\mu_i \otimes p_i + p_i)\right) = 0.$$

If the mean model and the contamination probability for each observation are correctly specified, this MCWGEE method provides a working estimate of regression parameters under certain assumptions. However, challenges remain in trying to ac-

curately estimate the contamination probability p_{it}. Through the exploration of different weighting schemes, we aim to find accurately estimated w_{it} that could reduce the bias of the parameter's estimations and improve the consistency of the MCWGEE approach.

3.2.3 Approaches for assigning weight in the MCWGEE approach

3.2.3.1 Subject specific weight

In our working framework, urines are collected every k days, and the time period for cocaine to be cleared from urine is h days ($h \leq k$) (Figure 2.1). We plan to divide the whole time period into multiple $k - day$ blocks, and calculate the summation of self-reported cocaine use of h days, $\sum_{t=k\times j-h+1}^{k\times j} Z_{it}$, for $j = 1, ..., m_i$, m_i is the number of urine measures for subject i. An indicator variable I_{ij} is defined as $I_{ij} = 0$ if $\sum_{t=k\times j-h+1}^{k\times j} Z_{it} = 0, I_{ij} = 1$ otherwise.

The difference between each urine test result U_{ij} and the indicator variable I_{ij} is $F_{ij} = |U_{ij} - I_{ij}|$, where $i = 1, ..., N; j = 1, ..., m_i$. We then use F_{ij}, the difference variable to detect contamination, in which case $F_{ij} = 0$ indicates that we fail to detect contamination, while $F_{ij} = 1$ suggests a contamination detection in the block. Since the weight is defined as the inverse of the contamination probability, $w_{it} = p_{it}$,

the subject specific weight is estimated as:

$$\hat{w}_i = \frac{\sum_{j=1}^{m_i} F_{ij}}{m_i}.$$

However, this approach assumes that the contamination probability for the first $k-h$ days in each time block is the same as the last h days, which may not be true. Moreover, even if we successfully detect contamination in a block, it is still challenging to locate the exact timepoints of contamination within the block. In some scenarios, the contamination probability may be underestimated.

3.2.3.2 Observation specific weight

We also consider a model based on the observation specific weight. In this approach, in order to model the contamination probability, we assume that the contamination indicator R_{it} depends on some covariates and can be modeled through logistic regression models.

As defined earlier, $R_{it} = 1$ if self-reported drug use is not the same as true drug use, otherwise $R_{it} = 0$. Assume

$$\log \frac{Pr\left(R_{it} = 1 | X_i, B_{it}, \theta\right)}{1 - Pr\left(R_{it} = 1 | X_i, B_{it}, \theta\right)} = \theta_0 + \theta_1 X_i + \theta_2 B_{it},$$

$$Pr\left(R_{it}=1|X_i,B_{it},\theta\right)=\frac{e^{\theta_0+\theta_1X_i+\theta_2B_{it}}}{1+e^{\theta_0+\theta_1X_i+\theta_2B_{it}}},$$

where X_i denotes a vector of time independent covariates, B_{it} denotes a vector of time dependent covariates, and θ represents a vector of the regression parameters.

To estimate the contamination probability using urine data, we fit a model using the difference between urine test result and the indicator variable, F_{ij} (which is calculated in the last section), time independent covariates, X_i, and a function of time dependent covariates for h days, B'_{ij}:

$$\log\frac{Pr\left(F_{ij}=1|X_i,B'_{ij},\theta'\right)}{1-Pr\left(F_{ij}=1|X_i,B'_{ij},\theta'\right)}=\theta'_0+\theta'_1X_i+\theta'_2B'_{ij},$$

where $i=1,...,N;j=1,...,m_i$, m_i is the number of urine measures for subject i, and $\theta'_0,\theta'_1,\theta'_2$ are the regression parameters. Estimations of $\theta'_0,\theta'_1,\theta'_2$, i.e., $\hat{\theta}'_0,\hat{\theta}'_1,\hat{\theta}'_2$ are used to model $\hat{Pr}(R_{it}=1)$.

$\hat{Pr}\left(R_{it}=1|X_i,B_{it},\hat{\theta}'\right)$ will be estimated by the following model,

$$\log\frac{\hat{Pr}\left(R_{it}=1|X_i,B_{it},\hat{\theta}'\right)}{1-\hat{Pr}\left(R_{it}=1|X_i,B_{it},\hat{\theta}'\right)}=\hat{\theta}'_0+\hat{\theta}'_1X_i+\hat{\theta}'_2B_{it},$$

where $i=1,...,N;t=1,...,T$.

And the observation specific weight is defined as:

$$\hat{w}_{it} = \hat{Pr}\left(R_{it} = 1 | X_i, B_{it}, \hat{\theta}'\right) = \frac{e^{\hat{\theta}'_0 + \hat{\theta}'_1 X_i + \hat{\theta}'_2 B_{it}}}{1 + e^{\hat{\theta}'_0 + \hat{\theta}'_1 X_i + \hat{\theta}'_2 B_{it}}},$$

where $i = 1, ..., N; t = 1, ..., T$.

The estimation of the regression parameters may not be asymptotically unbiased when the estimation of the weight, the contamination probability, is biased. In the next section, we address the asymptotic normality of estimators from MCWGEE approach under the true value of contamination probability. And in section 3.4, we examine the asymptotic bias of estimators based on the MCWGEE approach for self-reported data when the estimation of the contamination probability is biased.

3.3 Asymptotic Properties of the Estimators

Under the assumptions that model means and drop-out process (MAR) are correctly specified, Robins et al. (1995) proposed that the WGEE yielded a consistent estimate of β, and $\sqrt{N}(\hat{\beta} - \beta)$ is asymptotically normally distributed with mean zero and estimated variance matrix:

$$\left(\sum_{i=1}^{N} \hat{D}'_i \hat{V}_i^{-1} \hat{W}_i \hat{D}_i\right)^{-1} \left(\sum_{i=1}^{N} \hat{G}_i \hat{G}'_i\right) \left(\sum_{i=1}^{N} \hat{D}'_i \hat{V}_i^{-1} \hat{W}_i \hat{D}_i\right)^{-1},$$

where

$$\hat{G}_i = \hat{U}_i - (\sum_{i=1}^{N} \hat{U}_i \hat{S}_i')(\sum_{i=1}^{N} \hat{S}_i \hat{S}_i')\hat{S}_i,$$

$$\hat{U}_i = \hat{D}_i' \hat{V}_i^{-1} \hat{W}_i \left(Z_i - \hat{\mu}_i\right),$$

and S_i is the score component for the ith individual from the drop-out model. The use of $\left(\sum_{i=1}^{N} \hat{G}_i \hat{G}_i'\right)$ instead of $\left(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{A}_i \hat{V}_i^{-1} \hat{D}_i\right)$ from the GEE approach in Chapter 2, adjusts for estimation of parameters in the drop-out model (Robins et al. 1995).

However, in our case, we use the WGEE model to correct report bias in the self-reported binary outcomes by assigning a larger weight to the observation with a smaller contamination probability, and a smaller weight to the observation with a larger contamination probability. Since our focus is not on missing data and we use a MCWGEE approach mainly to include the contamination probability into the estimating equation, our study differs from Robins et al.' work.

In this section, we study the asymptotic properties of $\hat{\beta}$, the solution of the MCWGEE of self-reported data by proving the existence, the weak consistency, and the asymptotic normality of the MCWGEE estimator $\hat{\beta}$ as sample size $N \to \infty$ and time periods for each subject T is bounded for all subjects. Our study has been built upon the framework of Yuan and Jennrich (1998) and results from Chapter 2.

The MCWGEE form of the self-reported data Z_i is:

$$\begin{aligned} U_N(\beta) &= \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} W_i \left(Z_i - \mu_i^*\right) \\ &= \sum_{i=1}^{N} (1 - 2p_i) \otimes X_i' A_i (A_i^{*\frac{1}{2}} C_i^*(\gamma) A_i^{*\frac{1}{2}})^{-1} W_i \left(Z_i - (\mu_i - 2\mu_i \otimes p_i + p_i)\right), \end{aligned}$$

where the correlation parameter γ the contamination probabilities p_i are assumed to be known, and $W_i = diag(\frac{1}{p_{i1}}, ..., \frac{1}{p_{iT}})$ is the weight matrix with weight defined as the inverse of the contamination probability of the same time point.

Here we aim to show that the solution $\hat{\beta}_N$ to $U_N(\beta) = 0$ is consistent,

$$\hat{\beta}_N \to \beta_0,$$

almost surely for the true value β_0, and $\hat{\beta}_N$ is approximately normally distributed as $N \to \infty$.

The Assumptions we need to prove the consistency and asymptotic normality of $\hat{\beta}_N$ are:

Assumption A. The subjects are independently sampled and there exists an upper bound $M < \infty$ such that the number of replicates $m_i < M$ for all subjects $i = 1, 2, ...$

Assumption B. There exists an upper bound $b < \infty$ such that $|X_i| < b$ for all subjects $i = 1, 2, ...$

Assumption C. It is assumed that $\frac{1}{N}\sum_{i=1}^{N} X_i X_i' \to B$ as $N \to \infty$, where B is a positive definite matrix.

Assumption A ensures that data from a finite number of subjects and do not dominate the parameter estimator. Assumption B ensures that the estimating functions $\frac{1}{N}U_N(\beta)$ and its first and higher-order derivatives with respect to beta are bounded. Assumption C means that for sufficiently large N, $\frac{1}{N}E(\frac{\partial}{\partial\beta'}U_N(\beta))$ will be positive definite, and there is no redundancy in the predictors.

Define the matrices

$$I_0^*(\beta) = \lim_{N\to\infty}\frac{1}{N}\frac{\partial}{\partial\beta'}U_N(\beta) = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}(D_i^{*\prime}V_i^{*-1}W_iD_i^*),$$

and

$$I_1^*(\beta) = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}(D_i^{*\prime}V_i^{*-1}W_iA_i^*W_iV_i^{*-1}D_i^*).$$

The existence of these limits is ensured by Assumption 2. Moreover, Assumption 3 ensures that $I_0^*(\beta)$ and $I_1^*(\beta)$ are positive definite.

The following Theorem shows that the MCWGEE is strongly consistent for β.

Theorem 1. *Under Assumptions A-C, with probability one there exist zeros* $\hat{\beta}_N$ *of* $U_N(\beta) = 0$ *such that* $\hat{\beta}_N \to \beta_0$ *as* $N \to \infty$.

In the Appendix, we demonstrate that $\frac{1}{N}U_N(\beta_0) \to 0$ a.s., as $N \to \infty$, and that $\frac{1}{N}\frac{\partial}{\partial \beta'}U_N(\beta)$ converges uniformly to a non-stochastic limit which is nonsingular at β_0. The results then follow from Theorem 2 of Yuan and Jennrich (1997).

The following Theorem shows that the MCWGEE estimator is approximately normally distributed for large N.

Theorem 2. *Under Assummptions A-C,* $\sqrt{N}\left(\hat{\beta} - \beta_0\right) \underrightarrow{L} N\left(0, I_0^{*-1}(\beta_0)I_1^*(\beta_0)I_0^{*-1}(\beta_0)\right)$, *as* $N \to \infty$.

The above result is proved in the appendix. The variance-covariance estimator of $\hat{\beta}$ can be estimated as:

$$\hat{I}_0^* = \sum_{i=1}^{N}(\hat{D}_i^{*\prime}\hat{V}_i^{*-1}W_i\hat{D}_i^*),$$

and

$$\hat{I}_1^* = \sum_{i=1}^{N}\hat{D}_i^{*\prime}\hat{V}_i^{*-1}W_i\hat{A}_i^*W_i\hat{V}_i^{*-1}\hat{D}_i^*.$$

Therefore, the asymptotic properties of $\hat{\beta}$ holds with the accurate estimation of the contamination probability. However, if the estimation of this probability is misspecified, $\hat{\beta}$ may not be asymptotically unbiased. In the next section, we explore the asymptotic bias of $\hat{\beta}$ from the MCWGEE approach for self-reported binary data when the estimation of the contamination probability has deviated from the true

value.

3.4 Bias of the Estimators with Report Bias

If the model means are correctly specified, then the MCGEE approach from chapter 2 may provide an unbiased estimation of β. However, when the time period for cocaine to be cleared from urine increases, it is difficult to obtain accurate estimation of the contamination probability, and the bias of the MCGEE model's estimators increase. Thus, to further reduce the potential bias of the estimators, we include a weight which is the inverse of the contamination probability to the model equation to compensate for the potential bias caused by contamination.

The MCWGEE form of the self-reported data Z_i is:

$$\begin{aligned} U^*_\beta(\beta) &= \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} W_i \left(Z_i - \mu_i^*\right) \\ &= \sum_{i=1}^{N} (1-2p_i) \otimes X_i' A_i (A_i^{*\frac{1}{2}} C_i^*(\gamma) A_i^{*\frac{1}{2}})^{-1} W_i \left(Z_i - (\mu_i - 2\mu_i \otimes p_i + p_i)\right). \end{aligned}$$

Since the weight is defined as the inverse of the contamination probability at each time point, the weight $\frac{1}{p_i}$ can be assigned using two methods, reference section 3.2.3.

One method is subject specific weight, i.e., a single weight has been assigned to each subject, and all the observations from the same subject receive the same weight.

From previous discussions, we estimated the contamination probability p_i for the ith subject using:

$$\hat{p}_i = \frac{\sum_{j=1}^{m} F_{ij}}{m_i},$$

where m_i is the number of urine tests for subject i, F_{ij} is the difference between urine test results U_{ij} and the indicator of the summation of self-reported binary outcomes of h days I_{ij}. The subject specific weight is defined as $\frac{1}{\hat{p}_i}$.

The other approach is observation specific weight, in which case, weight might be different for each observation. We assume that the contamination indicator R_{it} depends on some covariates and can be modeled through logistic regression models.

From section 3.2.3.2, $\hat{p}_{it} = \hat{Pr}\left(R_{it} = 1|X_i, B_{it}, \hat{\theta}'\right)$ is estimated by the following model,

$$\log \frac{\hat{Pr}\left(R_{it} = 1|X_i, B_{it}, \hat{\theta}'\right)}{1 - \hat{Pr}\left(R_{it} = 1|X_i, B_{it}, \hat{\theta}'\right)} = \hat{\theta}'_0 + \hat{\theta}'_1 X_i + \hat{\theta}'_2 B_{it},$$

so

$$\hat{p}_{it} = \frac{e^{\hat{\theta}'_0 + \hat{\theta}'_1 X_i + \hat{\theta}'_2 B_{it}}}{1 + e^{\hat{\theta}'_0 + \hat{\theta}'_1 X_i + \hat{\theta}'_2 B_{it}}},$$

where $i = 1, ..., N; t = 1, ..., T$; X_i denotes a vector of time independent covariates; B_{it} denotes a vector of time dependent covariates. The observation specific weight is defined as $\frac{1}{p_{it}}$.

Assume the marginal mean of the true drug use for the ith subject is not equal to 0.5

for each time points, $\mu_i = (\mu_{i1}, ...\mu_{iT})' \neq (0.5, ..., 0.5)'$, the contamination probability is not equal to 0.5 for each time points, $p_i = (p_{i1}, ...p_{iT})' \neq (0.5, ..., 0.5)'$, and the estimated contamination probability is also not equal to 0.5 for every time points, $\hat{p}_i = (p_{i1}, ...p_{iT})' \neq (0.5, ..., 0.5)'$.

If we replace p_i by $\hat{p}_i$ in the estimating equation, then $E_{\beta_0}(U^*_\beta(\beta))$ may not be equal to 0.

$$\begin{aligned} E_{\beta_0}(U^*_\beta(\beta)) &= \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} W_i \left(E(Z_i) - \mu_i^*\right) \\ &= \sum_{i=1}^{N} (1 - 2\hat{p}_i) X_i' A_i (A_i^{*\frac{1}{2}} C_i^*(\gamma) A_i^{*\frac{1}{2}})^{-1} W_i (1 - 2\mu_i)(p_i - \hat{p}_i). \end{aligned}$$

The above equation is only guaranteed to equal 0 when $\hat{p}_i = p_i$, i.e., when the contamination probability has been correctly estimated, we can have unbiased estimators.

From section 3.3, we have

$$E_{\beta_0}(U^*_\beta(\beta)) = \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} W_i \left(E(Z_i) - \mu_i^*\right) = 0,$$

$$\hat{\beta}_N \to \beta_0,$$

as $N \to \infty$, where β_0 is the true value.

However, with $\hat{p}_i \neq p_i$,

$$E_{\beta_0}(U^*_\beta(\beta)) = \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} W_i \left(E(Z_i) - \mu_i^*\right) \neq 0.$$

Instead,

$$E_{\beta_0}(U^*_\beta(\beta^*)) = 0,$$

$$\hat{\beta}_N \to \beta^*,$$

as $N \to \infty$.

To estimate the asymptotic bias of $\hat{\beta}$ is equivalent to the difference of $\beta^* - \beta_0$. From Chapter 2, we observe that when the difference between $\hat{p}_i$ and p_i increases, the difference between $E(Z_i)$ and μ_i^* increases, and subsequently the bias of $\hat{\beta}$ increases. By adding the inverse of contamination probability as the weight function, the observations with higher contamination probability have lower weight in the equations. Therefore, we can further reduce the bias of $\hat{\beta}$ by assigning a lower weight to outcomes with higher contamination probability.

Since the above equation does not have a closed form solution of β, we may also borrow the idea from Rotnitzky and Wypij (1992): for any fixed β, the estimating equation, a function of (Z_i, R_i, X_i), has expectation given by the sum of all the possible situations times their respected probabilities. Then instead of solving for β in the above equations, we can simply consider an artificial sample comprised of one

observation for each possible combinations of Z_i, R_i, and X_i, which are weighted by their specific probabilities.

3.5 Simulations

3.5.1 Data generation

Two groups (Treatment and control) are considered in the data generation. In each group, there are $N/2$ subjects whose outcomes are repeatedly measured at T time points. True drug use data is generated as:

$$Pr\left(Y_{it} = 1 | X_i, \beta\right) = \mu_{it},$$

$$\log \frac{\mu_{it}}{1 - \mu_{it}} = X_i\beta = \beta_0 + \beta_1 X_i + \sigma_i,$$

for $i = 1, ..., N$, $t = 1, ..., T$, where X_i is the treatment indicator, $X_i = 1$ denotes the individual is in a treatment group, $X_i = 0$ denotes the individual is in a control group, and σ_i is a random effect variable following normal distribution with mean zero and a common variance $v = 0.04$.

Urine data is generated based on the true drug use data, for the ith patient, we use U_{ij} to denote urine test result, where $j = 1, ..., m_i$. m_i is the number of urine tests for subject i. Based on the notation used previously, the self-reported data can be

written as:

$$Z_{it} = Y_{it}\left(1 - R_{it}\right) + \left(1 - Y_{it}\right) R_{it}.$$

Contamination indicator R_{it} is generated using two methods to assess different weight estimation using the MCWGEE approach.

To obtain the subject specific weight of the MCWGEE model, R_{it} is generated by a relatively simple contamination probability assumption, which is not model based. Assuming there is p_1 probability of contamination among all subjects at one or several time points, and within these subjects there is p_2 probability that they report false drug use at each time point. Each observation is independent. Thus, the overall contamination probability p equals:

$$p = p_1 \times p_2.$$

And it can be estimated as:

$$\hat{p} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} R_{it}}{N \times T}.$$

For each contaminated subjects, the contamination probability at each time point can be estimated as:

$$\hat{p}_i = \frac{1}{T} \sum_{t=1}^{T} R_{it},$$

the probability of how many subjects have been contaminated has the estimated

form:

$$\hat{p}_1 = \frac{1}{N}\sum_{i=1}^{N} I_i \left(\sum_{t=1}^{T} R_{it} \geq 1 \right),$$

where I_i is the indicator, $I_i = 1$ if $\sum_{t=1}^{T} R_{it} \geq 1$; otherwise $I_i = 0$.

To estimate the effect of observation specific weight of the MCWGEE approach, R_{it} is generated by a model based on time dependent covariates B_{it} and time independent covariates X_i. Where X_i is the treatment effect, B_{it} is the Buprenorphone bottle open data, σ_i is a random effect variable following normal distribution with mean zero and a common variance $v = 0.04$. In our simulation, we assume the indicator variable R_{it} follows:

$$\log \frac{Pr\left(R_{it} = 1 | X_i, B_{it}, \theta\right)}{1 - Pr\left(R_{it} = 1 | X_i, B_{it}, \theta\right)} = \theta_0 + \theta_1 X_i + \theta_2 B_{it} + \sigma_i.$$

Bias of estimators corresponding to each generation method have been assessed under different conditions, such as: contamination probabilities, different time periods for cocaine to be cleared from urine, study time periods, and sample sizes.

All data simulations and analysis are carried out using the R software, and 1000 replications are performed for each run to obtain reliable results.

3.5.2 Simulation results

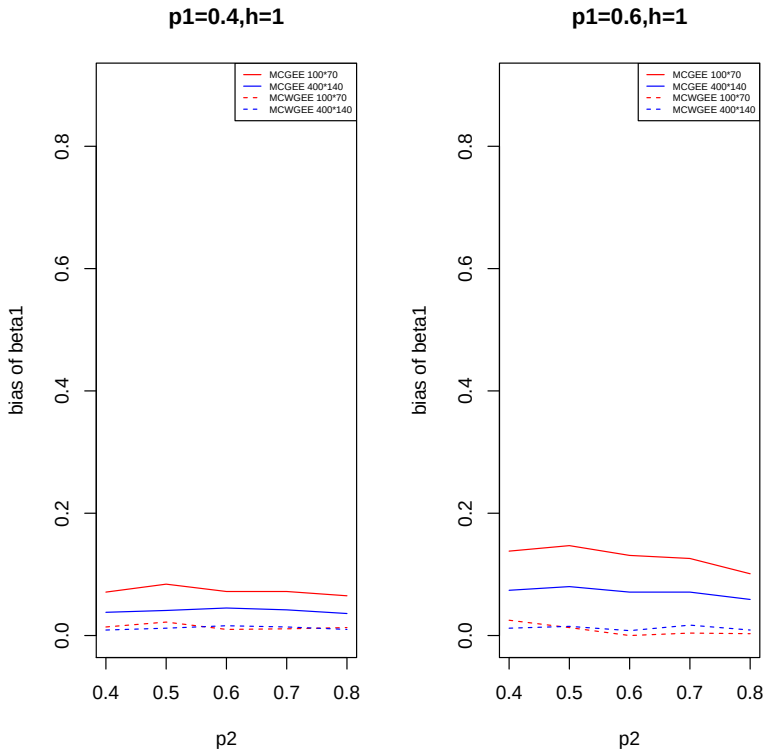
In this section, we compare the bias and standard error of the estimators of the WGEE approach and the MCWGEE approach for various situations under simulation. We assume that urines are collected every 7 days ($k = 7$). We generate $N/2$ subjects in the treatment and the control group, respectively, and whose outcomes are repeatedly measured at T time points. The true value for the intercept β_0 is 0.3, and the true value for the treatment effect β_1 is 1.2.

First, we consider the case where R_{it} is generated by a relatively simple contamination probability assumption to obtain a subject specific weight. As we discussed earlier, the bias of the MCWGEE approach of self-reported data exist when the estimated contamination probability is not the same as the true value ($\hat{p} \neq p$). Several combinations for different sample size ($N = 100$, $N = 400$), time periods for each individual's measurement ($T = 70$, $T = 140$), contamination probabilities ($p_1 = 0.4, 0.6$; $p_2 = 0.4, 0.8$), and different time periods for cocaine to be cleared from urine ($h = 1$, $h = 4$) have been considered.

Table 3.1 provides robust results with a small bias in the parameters' estimation for both WGEE and MCWGEE approach. The models performed well for each sample size, time period, and contamination probability. And the bias of the MCWGEE estimators are lower than the bias of the WGEE estimators. Compared with the MCGEE approach presented in Chapter 2, the MCWGEE has a smaller bias and higher coverage probability of 95% CI for each scenario. Moreover, the standard

error of the parameters decreases as the sample size and time period increase. The bias of β_1 for MCGEE and MCWGEE are also presented in Figure 3.1.

Figure 3.1: Bias of β_1 of MCGEE and MCWGEE when $h = 1$



After increasing the time period for cocaine to be cleared from urine from 1 day to 4 days (k=7, h=4), we observe that the bias for both models have increased when compared to $h = 1$. The bias of parameter estimates of the MCWGEE approach is still lower than that of the WGEE approach. Under multiple scenarios, their performances on bias correction decline significantly because the probability to detect the contamination measurement decreases as the time period for cocaine to be cleared

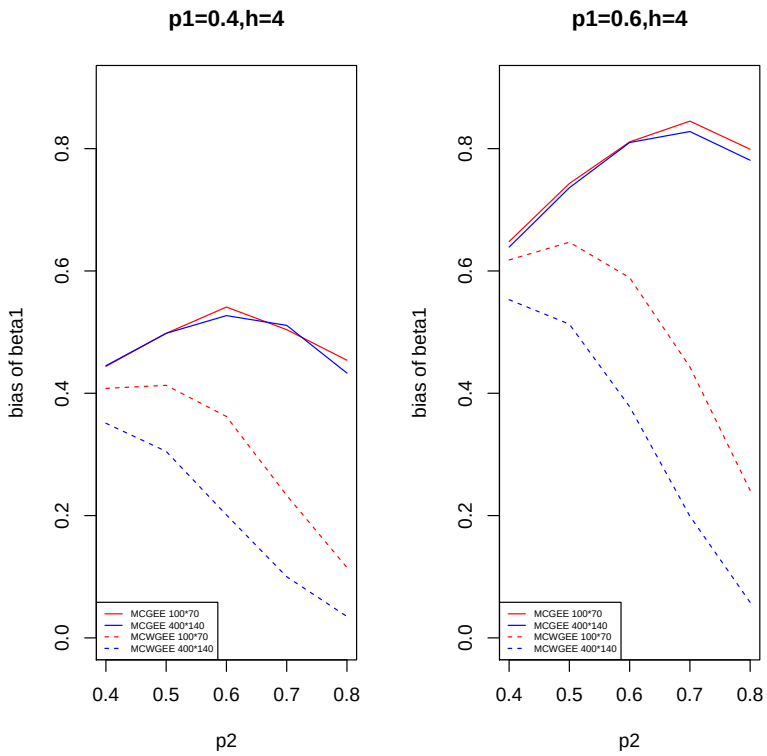
Table 3.1: Parameters' estimation, standard error (S.E.), and coverage probability of 95% CI (CP%) of the Mean Corrected GEE approach, subject specific WGEE approach, and the Mean Corrected subject specific WGEE approach (k=7, h=1).

Effect	N	T	p_1	p_2	MCGEE	CP%	WGEE	CP%	MCWGEE	CP%
$\beta_0(S.E.)$	100	70	0.4	0.4	0.279(0.05)	94.0	0.289(0.06)	93.2	0.293(0.06)	95.2
				0.8	0.291(0.05)	94.6	0.296(0.05)	92.4	0.299(0.06)	95.2
			0.6	0.4	0.262(0.06)	91.8	0.282(0.07)	90.6	0.290(0.07)	95.0
				0.8	0.283(0.06)	94.2	0.286(0.07)	88.4	0.296(0.08)	94.2
		140	0.4	0.4	0.292(0.05)	94.6	0.293(0.04)	95.2	0.299(0.05)	94.6
				0.8	0.293(0.04)	94.6	0.293(0.04)	91.6	0.297(0.05)	94.2
			0.6	0.4	0.274(0.05)	92.4	0.281(0.05)	92.6	0.290(0.06)	95.0
				0.8	0.292(0.04)	95.0	0.293(0.05)	93.4	0.302(0.06)	94.6
	400	70	0.4	0.4	0.283(0.03)	90.4	0.293(0.02)	93.2	0.297(0.03)	94.8
				0.8	0.289(0.02)	92.0	0.293(0.03)	93.6	0.297(0.03)	95.6
			0.6	0.4	0.267(0.03)	82.6	0.287(0.03)	92.2	0.296(0.04)	94.6
				0.8	0.282(0.03)	89.8	0.288(0.03)	93.4	0.297(0.03)	94.8
		140	0.4	0.4	0.288(0.02)	91.2	0.291(0.02)	91.8	0.295(0.02)	94.2
				0.8	0.293(0.02)	92.2	0.293(0.02)	92.0	0.297(0.02)	94.2
			0.6	0.4	0.281(0.03)	89.0	0.289(0.02)	93.6	0.297(0.03)	95.4
				0.8	0.290(0.02)	92.3	0.288(0.03)	94.0	0.297(0.03)	95.7
$\beta_1(S.E.)$	100	70	0.4	0.4	1.129(0.09)	87.6	1.162(0.09)	92.6	1.186(0.09)	94.2
				0.8	1.135(0.08)	86.4	1.163(0.08)	93.2	1.187(0.09)	94.2
			0.6	0.4	1.062(0.11)	74.2	1.125(0.10)	89.4	1.175(0.11)	94.2
				0.8	1.099(0.09)	79.0	1.144(0.10)	87.4	1.197(0.11)	95.4
		140	0.4	0.4	1.162(0.07)	92.0	1.167(0.07)	93.2	1.191(0.07)	95.4
				0.8	1.161(0.07)	91.2	1.164(0.07)	90.6	1.189(0.08)	94.4
			0.6	0.4	1.131(0.09)	88.6	1.142(0.08)	89.0	1.195(0.09)	93.6
				0.8	1.141(0.07)	88.0	1.133(0.08)	87.0	1.188(0.09)	95.2
	400	70	0.4	0.4	1.128(0.05)	65.4	1.160(0.04)	84.6	1.184(0.04)	93.6
				0.8	1.137(0.04)	68.0	1.166(0.04)	89.2	1.189(0.04)	94.2
			0.6	0.4	1.055(0.06)	25.4	1.121(0.05)	65.8	1.171(0.06)	91.6
				0.8	1.099(0.04)	35.6	1.140(0.05)	77.4	1.192(0.05)	95.6
		140	0.4	0.4	1.162(0.04)	79.6	1.168(0.04)	86.4	1.191(0.03)	94.4
				0.8	1.164(0.03)	82.6	1.167(0.04)	86.8	1.190(0.03)	93.4
			0.6	0.4	1.126(0.04)	59.4	1.137(0.04)	68.0	1.188(0.04)	94.4
				0.8	1.141(0.04)	62.5	1.139(0.04)	69.9	1.191(0.04)	95.7

Table 3.2: Parameters' estimation, standard error (S.E.), and coverage probability of 95% CI (CP%) of the Mean Corrected GEE approach, subject specific WGEE approach, and the Mean Corrected subject specific WGEE approach(k=7, h=4).

Effect	N	T	p_1	p_2	MCGEE	CP%	WGEE	CP%	MCWGEE	CP%
$\beta_0(S.E.)$	100	70	0.4	0.4	0.225(0.04)	58.4	0.254(0.04)	84.4	0.256(0.05)	85.4
				0.8	0.157(0.06)	26.2	0.227(0.05)	69.4	0.231(0.06)	76.6
			0.6	0.4	0.188(0.04)	30.0	0.227(0.05)	65.2	0.230(0.05)	72.8
				0.8	0.058(0.06)	2.2	0.165(0.05)	42.0	0.173(0.07)	53.8
		140	0.4	0.4	0.225(0.04)	47.6	0.266(0.04)	86.2	0.270(0.04)	89.0
				0.8	0.150(0.05)	13.0	0.254(0.04)	82.6	0.261(0.05)	87.8
			0.6	0.4	0.183(0.04)	10.2	0.241(0.04)	69.2	0.247(0.05)	78.8
				0.8	0.058(0.05)	0.6	0.218(0.05)	64.6	0.230(0.06)	78.0
	400	70	0.4	0.4	0.228(0.02)	8.6	0.255(0.02)	51.4	0.257(0.02)	53.4
				0.8	0.156(0.03)	0.0	0.228(0.02)	25.2	0.232(0.03)	30.4
			0.6	0.4	0.187(0.02)	0.0	0.225(0.02)	13.8	0.229(0.03)	21.8
				0.8	0.059(0.03)	0.0	0.164(0.02)	1.4	0.171(0.03)	1.8
		140	0.4	0.4	0.227(0.02)	4.6	0.268(0.02)	65.4	0.272(0.02)	75.4
				0.8	0.153(0.03)	0.0	0.258(0.02)	58.6	0.265(0.02)	70.4
			0.6	0.4	0.183(0.02)	0.0	0.242(0.02)	27.2	0.248(0.02)	38.4
				0.8	0.057(0.03)	0.0	0.216(0.02)	11.0	0.228(0.03)	23.6
$\beta_1(S.E.)$	100	70	0.4	0.4	0.756(0.10)	1.6	0.788(0.10)	2.8	0.792(0.11)	3.0
				0.8	0.746(0.16)	16.0	1.040(0.12)	79.0	1.085(0.13)	86.6
			0.6	0.4	0.552(0.10)	0.0	0.579(0.10)	0.0	0.582(0.11)	0.0
				0.8	0.401(0.19)	2.6	0.884(0.16)	51.8	0.959(0.19)	78.2
		140	0.4	0.4	0.759(0.09)	0.6	0.844(0.10)	4.6	0.853(0.10)	7.4
				0.8	0.782(0.15)	20.0	1.120(0.08)	83.8	1.175(0.08)	93.0
			0.6	0.4	0.566(0.09)	0.0	0.648(0.10)	0.0	0.656(0.11)	0.0
				0.8	0.406(0.19)	1.4	1.023(0.11)	63.8	1.137(0.10)	90.0
	400	70	0.4	0.4	0.750(0.05)	0.0	0.782(0.05)	0.0	0.786(0.05)	0.0
				0.8	0.743(0.08)	0.0	1.039(0.06)	24.0	1.082(0.06)	54.4
			0.6	0.4	0.554(0.05)	0.0	0.579(0.04)	0.0	0.582(0.06)	0.0
				0.8	0.412(0.09)	0.0	0.893(0.08)	2.6	0.965(0.09)	25.0
		140	0.4	0.4	0.755(0.05)	0.0	0.840(0.05)	0.0	0.849(0.05)	0.0
				0.8	0.767(0.07)	0.0	1.110(0.04)	37.4	1.165(0.04)	87.4
			0.6	0.4	0.561(0.05)	0.0	0.640(0.05)	0.0	0.647(0.05)	0.0
				0.8	0.419(0.09)	0.0	1.033(0.05)	8.0	1.142(0.05)	82.2

Figure 3.2: Bias of β_1 of MCGEE and MCWGEE when $h = 4$



from urine increases. When p_2 increases, MCWGEE approach results in less biased results compared to the MCGEE model. One possible explanation may be that with higher p_2, it is easier to detect contamination, and subjects with detected contaminations have been assigned to a relatively lower weight subsequently (Table 3.2, Figure 3.2).

Second, we consider the case where R_{it} is generated by a model based on the bottle

Table 3.3: Parameter value for model based generation of R_{it}.

$\bar{p}$	θ_0	θ_1	θ_2
0.1	0.05	0.5	-8.0
0.3	0.8	4.5	-3.0
0.5	1.4	5.5	-2.0

open data B_{it} and the treatment indicator X_i to obtain an observation specific weight:

$$\log \frac{Pr\left(R_{it}=1|X_i, B_{it}, \theta\right)}{1-Pr\left(R_{it}=1|X_i, B_{it}, \theta\right)} = \theta_0 + \theta_1 X_i + \theta_2 B_{it} + \sigma_i.$$

Several contamination probabilities estimated by the true R_{it} have been evaluated. The θ values we used to generate R_{it} for different mean of contamination probabilities are in Table 3.3.

Table 3.4 shows that the use of observation specific WGEE may result in biased estimates when contamination probability is high ($\bar{p} = 0.5$). When contamination probability is low ($\bar{p} = 0.1$), both methods provide less biased estimates. The bias of observation specific WGEE approach increases as the contamination probability increases. On the other hand, this bias remains relatively small for the observation specific MCWGEE model regardless of the change in the contamination probability. Moreover, the standard error of parameters for both models decreases as the sample size and the time period for each subject increase.

After increasing the time period for cocaine to be cleared from urine from 1 day to 4

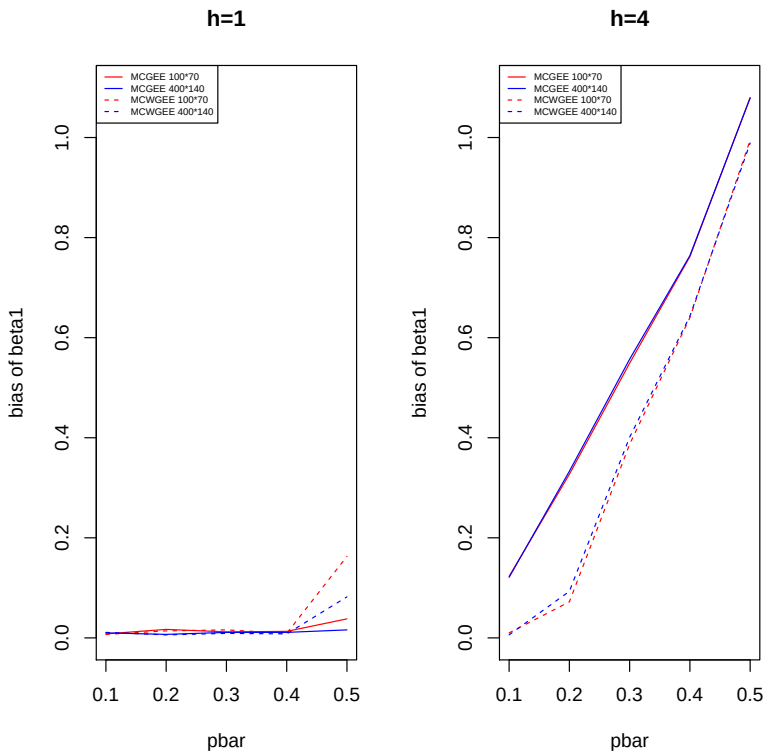
Table 3.4: Parameters' estimation, standard error (S.E.), and coverage probability of 95% CI (CP%) of the MCGEE approach, observation specific WGEE approach, and the MCWGEE approach (k=7, h=1).

Effect	N	T	$\bar{p}$	MCGEE	CP%	WGEE	CP%	MCWGEE	CP%
$\beta_0(S.E.)$	100	70	0.1	0.293(0.04)	85.8	0.294(0.05)	92.8	0.295(0.05)	92.8
			0.3	0.293(0.04)	88.6	0.220(0.05)	54.0	0.296(0.04)	87.0
			0.5	0.298(0.06)	90.2	0.032(0.07)	4.2	0.273(0.06)	67.6
		140	0.1	0.299(0.03)	82.6	0.299(0.04)	95.6	0.300(0.05)	95.6
			0.3	0.294(0.03)	84.0	0.213(0.04)	37.0	0.289(0.03)	87.8
			0.5	0.296(0.04)	83.4	0.023(0.06)	3.4	0.259(0.04)	68.6
	400	70	0.1	0.299(0.02)	89.0	0.298(0.02)	95.8	0.298(0.02)	96.0
			0.3	0.298(0.02)	89.0	0.221(0.02)	4.6	0.299(0.02)	92.2
			0.5	0.299(0.03)	89.6	0.036(0.03)	2.4	0.281(0.03)	70.4
		140	0.1	0.297(0.01)	78.0	0.297(0.02)	93.4	0.297(0.02)	93.0
			0.3	0.298(0.02)	84.4	0.220(0.02)	0.6	0.298(0.02)	88.4
			0.5	0.297(0.02)	85.2	0.037(0.02)	2.6	0.278(0.02)	68.0
$\beta_1(S.E.)$	100	70	0.1	1.192(0.09)	95.4	1.192(0.08)	94.6	1.193(0.07)	94.8
			0.3	1.188(0.13)	91.6	0.788(0.07)	0.2	1.184(0.06)	84.6
			0.5	1.162(0.10)	90.6	0.080(0.12)	2.8	1.037(0.09)	71.8
		140	0.1	1.187(0.08)	94.2	1.185(0.06)	94.6	1.187(0.08)	94.6
			0.3	1.192(0.10)	93.8	0.790(0.07)	0.4	1.185(0.06)	89.2
			0.5	1.180(0.11)	76.8	0.127(0.11)	2.4	1.115(0.12)	72.6
	400	70	0.1	1.188(0.04)	94.8	1.186(0.04)	92.0	1.188(0.05)	92.4
			0.3	1.187(0.06)	88.2	0.796(0.04)	0.0	1.188(0.03)	80.6
			0.5	1.178(0.05)	87.0	0.114(0.05)	0.2	1.111(0.05)	63.6
		140	0.1	1.190(0.05)	94.8	1.188(0.03)	92.8	1.189(0.03)	93.8
			0.3	1.189(0.04)	91.8	0.797(0.03)	0.0	1.191(0.04)	85.0
			0.5	1.184(0.04)	78.6	0.123(0.05)	0.0	1.118(0.04)	74.4

Table 3.5: Parameters' estimation standard error (S.E.), and coverage probability of 95% CI (CP%) of the MCGEE approach, the observation specific WGEE approach, and the MCWGEE approach (k=7, h=4).

Effect	N	T	$\bar{p}$	MCGEE	CP%	WGEE	(%)	MCWGEE	(%)
$\beta_0(S.E.)$	100	70	0.1	0.274(0.04)	79.4	0.298(0.05)	94.8	0.300(0.04)	94.8
			0.3	0.177(0.04)	17.8	0.199(0.04)	35.0	0.221(0.06)	56.0
			0.5	0.039(0.04)	0.0	0.047(0.04)	0.0	0.065(0.04)	0.0
		140	0.1	0.271(0.03)	67.8	0.295(0.04)	94.2	0.297(0.03)	93.6
			0.3	0.182(0.03)	6.8	0.198(0.04)	24.6	0.220(0.05)	41.2
			0.5	0.034(0.03)	0.0	0.043(0.03)	0.0	0.061(0.03)	0.0
	400	70	0.1	0.273(0.02)	63.6	0.298(0.02)	94.8	0.299(0.02)	94.0
			0.3	0.183(0.02)	0.0	0.203(0.02)	0.8	0.224(0.03)	7.8
			0.5	0.037(0.02)	0.0	0.050(0.02)	0.0	0.064(0.02)	0.0
		140	0.1	0.269(0.01)	37.8	0.294(0.02)	93.6	0.296(0.01)	94.6
			0.3	0.179(0.02)	0.0	0.197(0.02)	0.0	0.219(0.02)	0.8
			0.5	0.036(0.02)	0.0	0.045(0.01)	0.0	0.063(0.02)	0.0
$\beta_1(S.E.)$	100	70	0.1	1.077(0.09)	80.4	1.179(0.07)	92.4	1.190(0.08)	93.2
			0.3	0.651(0.11)	0.2	0.713(0.08)	0.0	0.813(0.10)	0.0
			0.5	0.120(0.10)	0.0	0.143(0.07)	0.0	0.205(0.08)	0.0
		140	0.1	1.075(0.09)	85.4	1.180(0.06)	93.6	1.192(0.06)	94.2
			0.3	0.637(0.08)	0.0	0.687(0.08)	0.0	0.791(0.07)	0.0
			0.5	0.124(0.08)	0.0	0.145(0.07)	0.0	0.208(0.08)	0.0
	400	70	0.1	1.082(0.06)	39.6	1.182(0.04)	89.6	1.193(0.03)	93.6
			0.3	0.648(0.06)	0.0	0.716(0.04)	0.0	0.817(0.05)	0.0
			0.5	0.122(0.04)	0.0	0.151(0.04)	0.0	0.215(0.04)	0.0
		140	0.1	1.079(0.04)	34.6	1.181(0.03)	90.0	1.194(0.02)	94.0
			0.3	0.643(0.04)	0.0	0.697(0.04)	0.0	0.799(0.04)	0.0
			0.5	0.121(0.04)	0.0	0.148(0.03)	0.0	0.211(0.05)	0.0

Figure 3.3: Bias of β_1 of MCGEE and MCWGEE



days ($h = 4$), both WGEE and MCWGEE perform poorly on bias correction when contamination probability is high ($\bar{p} = 0.5$). Under lower contamination probability ($\bar{p} = 0.1$), both of these methods report less biased estimators. In addition, the bias of the MCWGEE estimators are lower than the WGEE estimators under each situation (Table 3.5). From Figure 3.3, the bias of β_1 of MCWGEE seems similar to that of MCGEE when $h = 1$ and $\bar{p}$ is relatively low, although both bias of these two methods increase significantly when $h = 4$, the bias of MCWGEE is generally lower

than MCGEE's bias (Figure 3.3).

3.6 SCU Data

From Chapter 2, there are a total of 140 patients in the Self-reported Cocaine use and Urine test (SCU) data, followed for a period of 5-6 months. After a 2-week induction and stabilization period, during which patients were treated by nurses 3 times per week with 16 mg buprenorphine daily, enrolled subjects were randomly assigned to the treatment or the control group. Both groups received buprenorphine, a substitute for cocaine use, which was stored in bottles. Buprenorphine was instructed to be taken once per day. If the bottle was opened on a specific day, the patient was regarded as adherent. The special MEMSCAP bottles can record the time when the bottle is opened.

The control group received physical management (PM), a 15-20 minutes session by Internal Medicine physicians with experiences as buprenorphine providers. Throughout the study period, sessions occurred weekly for the first two weeks, every two weeks for the next four weeks, and then monthly. The treatment group received PM plus cognitive behavioral therapy (CBT). CBT is a counseling intervention that has demonstrated efficacy in treating a variety of psychiatric conditions and cocaine dependences. CBT was provided by masters- and doctoral-level clinicians who were trained with a manual adapted from a guidance for the use of CBT for cocaine de-

pendence (Carroll 1998). The main components of counseling focused on performing a functional analysis of behavior, promoting behavioral activation, identifying and coping with drug cravings, enhancing drug refusal skills and decision makings about high risk situations, and improving problem solving skills (Fiellin et al. 2013).

The study's major outcomes include (1) self-reported daily illicit drug uses which were reported during the weekly PM sessions, and (2) weekly urine test results. Self-reported daily illicit drug uses variables include: cocaine use, marijuana use, alcohol use, bup use, and prescopioid/heroin/opium/other opiate use. Urine test variables include: cocaine, benzo, THC, and opiate/methadone/oxycontin. Overall, there are five self-reported variables and four urine test variables. Our main interest in the motivating example is the cocaine use, including the self-reported cocaine use and the weekly cocaine urine test. After statistical methods being developed for the illustrative example on these variables, they may also be extended to other substances' uses.

As defined in the previous sections, let Z_{it} denote the self-reported daily cocaine use, X_i denote the treatment effect, with $X_i = 1$ indicates patient in the treatment group. We first conduct the subject specific WGEE model using self-reported data with the logistic link:

$$\log \frac{Pr\left(Z_{it} = 1 | X_i, \beta\right)}{1 - Pr\left(Z_{it} = 1 | X_i, \beta\right)} = \beta_0 + \beta_1 X_i.$$

We assume the contamination indicator R_{it} does not depend on the MEMSCAP bottle open data, and the time period for cocaine to be cleared from urine is 1

day ($h = 1$). We use urine test results to estimate the contamination probability, and assign this subject specific weight as the inverse of the contamination probability.

Table 3.6: Results of subject specific WGEE of cocaine use (h=1)

	Estimate	S.E.	P-value
β_0	-2.41	0.11	<0.0001
β_1	0.26	0.18	0.14

Table 3.6 indicates that the p-value for effect of treatment is 0.14, suggesting that there is no significant treatment effect.

Table 3.7: Results of subject specific WGEE approach of cocaine use (h=4)

	Estimate	S.E.	P-value
β_0	-2.23	0.15	<0.0001
β_1	0.28	0.21	0.19

After increasing the time period for cocaine to be cleared from urine to 4 days ($h = 4$), the effect of treatment increases slightly, there is no significant change in the treatment effect (Table 3.7).

Table 3.8: Results of subject specific MCWGEE approach of cocaine use (h=1)

	Estimate	S.E.	P-value
β_0	-3.16	0.11	<0.0001
β_1	0.31	0.17	0.07

After using contamination probability to correct the mean of self-reported data, we then apply the subject specific MCWGEE approach to estimate the effect of CBT.

Table 3.9: Results of subject specific MCWGEE approach of cocaine use (h=4)

	Estimate	S.E.	P-value
β_0	-2.45	0.12	<0.0001
β_1	0.35	0.18	0.05

Assuming the time period for cocaine to be cleared from urine is 1 day ($h = 1$), we can clearly see that the estimation of treatment effect has increased, the p-value for effect of treatment is 0.07 (Table 3.8). When $h = 4$, the effect of treatment also increases, the p-value for this effect has decreased to 0.05 (Table 3.9).

Table 3.10: Results of observation specific WGEE approach of cocaine use (h=1)

	Estimate	S.E.	P-value
β_0	-2.14	0.11	<0.0001
β_1	0.24	0.16	0.14

Table 3.11: Results of observation specific WGEE approach of cocaine use (h=4)

	Estimate	S.E.	P-value
β_0	-2.13	0.11	<0.0001
β_1	0.25	0.16	0.13

Second, we consider the case where the contamination indicator R_{it} depends on the MEMSCAP bottle open data, and build the observation specific weight as the inverse of the contamination probability at each time point. The effect of CBT hasn't changed much when compared to the subject specific WGEE (Table 3.10, Table 3.11).

We then use contamination probability to correct the marginal mean of self-reported data, and apply the subject specific MCWGEE model to estimate the effect of CBT. Under both circumstances when $h = 1$ and $h = 4$, the estimation of treatment effect increases, and the p-value for this effect has been decreases (Table 3.12, Table 3.13).

Table 3.12: Results of observation specific MCWGEE approach of cocaine use (h=1)

	Estimate	S.E.	P-value
β_0	-2.29	0.11	<0.0001
β_1	0.28	0.15	0.06

Table 3.13: Results of observation specific MCWGEE approach of cocaine use (h=4)

	Estimate	S.E.	P-value
β_0	-2.34	0.10	<0.0001
β_1	0.28	0.14	0.05

3.7 Discussion and Conclusion

In this chapter, we present the Mean Corrected subject specific WGEE approach and the Mean Corrected observation specific WGEE approach to analyze longitudinal binary self-reported outcomes with report bias. From Chapter 2, when the marginal mean of the self-reported results Z_{it} has been correctly specified, the MCGEE approach yields consistent estimates of the parameters. However, these properties rely on the validity of the estimation of the contamination probability. Estimators may not be asymptotically unbiased when the contamination probability is misspecified, i.e., the marginal means have not been correctly specified. Since it is more difficult

to estimate the contamination probability when the time period for cocaine to be cleared from urine is longer than one day, the bias of the estimators tends to increase as the time period for cocaine to be cleared from urine increases. Thus, we add a weight, which is the inverse of the contamination probability into the equation and build a MCWGEE approach to further control the parameters' bias.

Traditionally, WGEE approach is often used under the Missing at Random (MAR) assumption for incomplete data with informative dropouts, and it models each subject's measurements by the inverse probability that a subject has each measurement observed. In our case, we use the MCWGEE approach to correct report bias in the self-reported binary outcomes by assigning a larger weight to the observation with a smaller contamination probability, and a smaller weight to the observation with a larger contamination probability. Since our focus is not on missing data and we use a MCWGEE approach mainly to include the contamination probability into the estimating equations, our study differs from the original WGEE approach of Robins et al (1995).

Comparing with the MCGEE approach in Chapter 2, we find that in most situations with combinations for different sample size, contamination probabilities, and time periods for cocaine to be cleared from urine, the bias of the MCWGEE estimators are generally lower than that of the MCGEE estimators, the coverage probability of 95% CI is higher for the MCWGEE approach compared to the MCGEE approach. MCWGEE performs better especially in the case when the time period for cocaine

to be cleared from urine is 4 days, since it is difficult to detect the time and location of contamination in each time block when time period for cocaine to be cleared from urine increases. Under this circumstance, it is difficult to correctly specify the marginal mean of the self-reported data in the estimating equation. The bias of the estimates of the MCGEE approach increases when the time period for cocaine to be cleared from urine increases. By adding an inverse of contamination probability as a weight to the MCGEE approach, the data with higher contamination probability yield a lower weight in the equation. Subsequently, by assigning a lower weight to the observations with higher contamination probability, the bias of the estimates of the MCWGEE approach decreases compared with the MCGEE approach.

Naturally, we can further reduce the bias of the estimators by assigning a lower weight to outcomes with higher contamination probability. We explored an extreme case in our preliminary analysis (results not included in the dissertation), which only includes subjects without contaminations. In other words, we built a WGEE model by assigning zero weight to the subjects with contamination. Based on our preliminary simulation results, this approach provides less biased estimates. However, the extreme approach may be considerably less efficient than the MCWGEE approach when the contamination probability is large. In such case, only a small proportion of the originally enrolled subjects are retained in the analysis and the statistical power is greatly reduced.

We have shown that the asymptotic properties of parameters from the MCWGEE

approach hold if the estimation of the contamination probability is accurate. However, it is more difficult to ensure the validity of this estimation when the time period for cocaine to be cleared from urine increases. Our simulation results indicate similar findings. When the time period for cocaine to be cleared from urine is 1 day ($h = 1$), both MCGEE and MCWGEE approaches result in less biased estimation (Figure 3.1). When the time period increases to 4 days ($h = 4$), the bias for both approaches increases.

Under the assumption that $h = 4$, for the subject specific MCWGEE approach, it performs better with a higher contamination probability at each time point within the contaminated subjects (p_2). This is because with higher p_2, it is easier to detect the contamination and assign a relatively lower weight to the identified subject. From Figure 3.2, we can observe that the bias of the estimates of treatment effect decreases as p_2 increases for MCWGEE approach. While the trend of the bias for MCGEE approach seems different. The effect of weight adjustment is more apparent when we have subjects with high frequencies of contamination in their responses. Therefore, the bias of the estimators decrease as the contamination probability for each subject increases for MCWGEE approach (Figure 3.2).

Under the assumption that contamination indicator depends on some covariates and is modeled through a logistic regression model, both the observation specific WGEE and observation specific MCWGEE provide satisfying results when the contamination probability is low ($\bar{p} = 0.1$). However, their performance on bias correction drop

significantly when the contamination probability is high ($\bar{p} = 0.5$). Compared with the MCGEE approach developed in Chapter 2, bias of the treatment estimator is lower for the MCWGEE approach especially when $h = 4$. Bias of the parameters' estimation of both MCGEE and MCWGEE approaches increases when the mean of overall contamination probability increases (Figure 3.3). We further explore the empirical evidence in our simulation scenarios of p_1 and p_2, and found that p_2 ranges from 0.2 to 0.8, while p_1 increases as $\bar{p}$ increases. Given this situation, future studies exploring wider ranges of p_2 is warranted.

After applying this MCWGEE approach to the SCU data, we notice that the estimated treatment effect has increased and the bias of the treatment effect is likely to be reduced when compared to the WGEE approach without mean correction. In comparison to the results of the MCGEE approach in Chapter 2, these estimates changed slightly. This finding reveals the feasibility of adding an inverse of contamination probability as a weight to the MCGEE approach.

3.8 Appendix

From section 3.3, $U_N(\beta) = \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} W_i (Z_i - \mu_i^*)$, define $\psi(Z_i; \beta) = D_i^{*\prime} V_i^{*-1} W_i (Z_i - \mu_i^*)$.
The Assumptions we need to prove the consistency and asymptotic normality of $\hat{\beta}_N$ are:

Assumption A. The subjects are independently sampled and there exists an upper

bound $M < \infty$ such that the number of replicates $m_i < M$ for all subjects $i = 1, 2, ...$

Assumption B. There exists an upper bound $b < \infty$ such that $|X_i| < b$ for all subjects $i = 1, 2, ...$

Assumption C. It is assumed that $\frac{1}{N}\sum_{i=1}^{N} X_i X_i' \to B$ as $N \to \infty$, where B is a positive definite matrix.

Under Assumption B and Assumption C, we define positive definite matrix

$$I_0^*(\beta) = \lim_{N\to\infty} \frac{1}{N} \frac{\partial}{\partial \beta'} U_N(\beta) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (D_i^{*\prime} V_i^{*-1} W_i D_i^*),$$

and

$$I_1^*(\beta) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (D_i^{*\prime} V_i^{*-1} W_i A_i^* W_i V_i^{*-1} D_i^*).$$

In order to prove the solution $\hat{\beta}_N$ of $U_N(\beta) = 0$ is consistent and asymptotic normally distributed for large N, we need to show that:

1. $\frac{1}{N} U_N(\beta_0) \to 0$ a.s., as $N \to \infty$.

2. $\frac{1}{N} \frac{\partial}{\partial \beta^T} U_N(\beta)$ converge uniformly to a nonstochastic limit which is nonsingular at β_0.

3. With probability one, $\psi(Z_i;\beta)$ are twice continuously differentiable with respect to $\beta \in B$, and $|\frac{\partial^2}{\partial\beta_j\partial\beta_k}\psi(Z_i;\beta)| < \infty$.

4. $|\psi(Z_i;\beta)| < \infty$, and $\frac{1}{\sqrt{N}}U_N(\beta_0) \underrightarrow{L} N(0, I_1^*(\beta_0))$.

Since our proof based on some assumptions and theorems from Yuan and Jennrich (1998), we verify their assumptions in our case in section 3.8.1, we prove the consistency of MCGEE estimator in section 3.8.2, and we show the asymptotic normality of MCGEE estimator in section 3.8.3.

3.8.1 Verifying the conditions

The conditions from Yuan and Jennrich (1998) we need to prove the consistency and asymptotic normality in Section 3.3 are:

1. $U_N(\beta_0) \to 0$ a.s., as $N \to \infty$.

2. There exists a neighborhood M of β_0 on which with probability one, all $U_N(\beta)$ are continuously differentiable and $\frac{\partial}{\partial\beta^T}U_N(\beta)$ converge uniformly to a nonstochastic limit which is nonsingular at β_0.

3.

$$\sqrt{N}U_N(\beta_0) \underrightarrow{L} N(0, I_1^*(\beta_0)),$$

as $N \to \infty$.

We verify condition 4 to imply condition 1 by Theorem 5 of Yuan and Jennrich (1998).

4. For each i, $\psi(Z_i;\beta_0)$ has mean zero and variance-covariance matrix K_i, such that

$$\frac{1}{N}\sum_{i=1}^{N} K_i \to K.$$

for some positive-definite matrix K.

For self-reported data Z_i, since $E(Z_i) = \mu_i^*$, then $E(\psi(Z_i;\beta_0)) = 0$.

$$var(\psi(Z_i;\beta_0)) = D_i^{*\prime} V_i^{*-1} W_i A_i^* W_i V_i^{*-1} D_i^*.$$

Since $|x_{it}| \leq b < \infty$ for all $i = 1, 2, ...$ and $t = 1, 2, ...T$, by assumption, we also add $\frac{1}{T}$ to the weight to ensure the weight is bounded, then

$$W_i = diag(\frac{1}{p_{i1}} + \frac{1}{T}, ..., \frac{1}{p_{iT}} + \frac{1}{T}) < \infty,$$

$$D_i^* = (1 - 2p_i) \otimes A_i X_i < \infty,$$

$$A_i^* = diag(var(Z_{i1}), ...var(Z_{iT})),$$

$$var(Z_{it}) = (1-2p_{it})^2 \frac{e^{x_{it}\beta}}{(1+e^{x_{it}\beta})^2} + p_{it}(1-p_{it}) < \infty,$$

$$V_i = A_i^{*\frac{1}{2}} C_i^*(\gamma) A_i^{*\frac{1}{2}} < \infty.$$

Thus, the variance-covariance matrix K_i of $\psi(\beta_0)$ follows

$$\frac{1}{N}\sum_{i=1}^{N} K_i \to K.$$

for some positive-definite matrix K. Condition 1 has been verified.

An equivalent approach to verify condition 2 is using the following conditions of Yuan and Jennrich (1998):

6. With probability one, $\psi(Z_i;\beta)$ are twice continuously differentiable with respect to $\beta \in B$.

7. For each $\beta \in B$,

$$\frac{1}{N}\sum_{i=1}^{N} E(\frac{\partial}{\partial \beta^T}\psi(Z_i;\beta)) \to I_0^*(\beta),$$

where $I_0^*(\beta) = \lim_{N\to\infty} \frac{1}{N}\frac{\partial}{\partial \beta'} U_N(\beta)$ is nonsingular and with probability one

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \beta_t}\psi(Z_i;\beta) \to I_0^*(\beta),$$

as $N \to \infty$.

8. For each i,

$$|\frac{\partial^2}{\partial\beta_j\partial\beta_k}\psi(Z_i;\beta)| \leq S,$$

for some upper bound $S < \infty$.

Yuan and Jennrich proved that under conditions 6, 7, and 8, condition 2 is satisfied.

To verify condition 6, we have

$$\begin{aligned}\frac{\partial}{\partial\beta}\psi(Z_i;\beta) =& \frac{\partial}{\partial\beta}(D_i^{*\prime}V_i^{*-1}W_i(Z_i-\mu_i^*)) \\ &= D_i^{*\prime}V_i^{*-1}W_iD_i^* + (\frac{\partial}{\partial\beta}D_i^*)'V_i^{*-1}W_i(Z_i-\mu_i^*) + D_i^{*\prime}(\frac{\partial}{\partial\beta}V_i^{*-1})W_i(Z_i-\mu_i^*).\end{aligned}$$

Since $E(Z_i) = \mu_i^*$, the last two terms in the expression above have expectation zero, so

$$E(\frac{\partial}{\partial\beta}\psi(Z_i;\beta)) = D_i^{*\prime}V_i^{*-1}W_iD_i^*.$$

Moreover,

$$\frac{\partial}{\partial\beta}D_i^* = (1-2p_i)\otimes(\frac{\partial}{\partial\beta}A_i)X_i,$$

where $A_i = diag(var(Y_{i1}), ..., var(Y_{iT}))$, and

$$\begin{aligned}\frac{\partial}{\partial\beta}\left(var(Y_{it})\right) =& \frac{\partial}{\partial\beta}\frac{e^{\beta' X_{it}}}{(1+e^{\beta' X_{it}})^2} \\ =& \frac{X_{it}e^{\beta' X_{it}}(1-e^{\beta' X_{it}})}{(1+e^{\beta' X_{it}})^3},\end{aligned}$$

$$\frac{\partial}{\partial\beta}A_i = diag\left(\frac{X_{i1}e^{\beta' X_{i1}}(1-e^{\beta' X_{i1}})}{(1+e^{\beta' X_{i1}})^3}, ..., \frac{X_{iT}e^{\beta' X_{iT}}(1-e^{\beta' X_{iT}})}{(1+e^{\beta' X_{iT}})^3}\right).$$

Since $|X_i| < b < \infty$, $0 \leq (1-2p_i) \leq 1$, $0 \leq \frac{e^{\beta' X_{it}}}{(1+e^{\beta' X_{it}})^2} \leq \frac{1}{4}$, and $0 \leq \frac{1}{1+e^{\beta' X_{it}}} \leq 1$,

$$\frac{\partial}{\partial\beta}D_i^* < \infty.$$

And,

$$\frac{\partial}{\partial\beta}V_i^{*-1} = -V_i^{*-1}(\frac{\partial}{\partial\beta}V_i^*)V_i^{*-1},$$

$$\frac{\partial}{\partial\beta}V_i^* = (\frac{\partial}{\partial\beta}A_i^{*\frac{1}{2}})C_i(\gamma)A_i^{*\frac{1}{2}} + A_i^{*\frac{1}{2}}C_i(\gamma)(\frac{\partial}{\partial\beta}A_i^{*\frac{1}{2}}).$$

$A_i^* = diag(var(Z_{i1}), ..., var(Z_{iT}))$, and

$$\begin{aligned}\frac{\partial}{\partial\beta}\left(var(Z_{it})\right) =& \frac{\partial}{\partial\beta}\left((1-2p_{it})^2\frac{e^{X_{it}\beta}}{(1+e^{X_{it}\beta})^2} + p_{it}(1-p_{it})\right) \\ =& (1-2p_{it})^2\frac{X_{it}e^{X_{it}\beta}(1-e^{X_{it}\beta})}{(1+e^{X_{it}\beta})^3}.\end{aligned}$$

Then

$$\begin{aligned}\frac{\partial}{\partial\beta}A_i^{*\frac{1}{2}} &= diag[\frac{\partial}{\partial\beta}\sqrt{(1-2p_{i1})^2\frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2}+p_{i1}(1-p_{i1})},...,\\ &\quad \frac{\partial}{\partial\beta}\sqrt{(1-2p_{iT})^2\frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2}+p_{iT}(1-p_{iT})}]\\ &= \frac{1}{2}diag[\frac{(1-2p_{i1})^2\frac{X_{i1}e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2}-2\frac{X_{i1}e^{2\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^3}}{\sqrt{(1-2p_{i1})^2\frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2}+p_{i1}(1-p_{i1})}},...,\\ &\quad \frac{(1-2p_{iT})^2\frac{X_{iT}e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2}-2\frac{X_{iT}e^{2\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^3}}{\sqrt{(1-2p_{iT})^2\frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2}+p_{iT}(1-p_{iT})}}]\\ &= \frac{1}{2}diag[\frac{(1-2p_{i1})^2\mu_{i1}(1-\mu_{i1})(1-2\mu_{i1})X_{i1}}{\sqrt{(1-2p_{i1})^2\frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2}+p_{i1}(1-p_{i1})}},...,\\ &\quad \frac{(1-2p_{iT})^2\mu_{iT}(1-\mu_{iT})(1-2\mu_{iT})X_{iT}}{\sqrt{(1-2p_{iT})^2\frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2}+p_{iT}(1-p_{iT})}}].\end{aligned}$$

Therefore, $\frac{\partial}{\partial\beta}A_i^{*\frac{1}{2}} < \infty$, $\frac{\partial}{\partial\beta}V_i^* < \infty$, $(\frac{\partial}{\partial\beta}D_i^*)'V_i^{*-1}W_i(Z_i-\mu_i^*)+D_i^{*\prime}(\frac{\partial}{\partial\beta}V_i^{*-1})W_i(Z_i-\mu_i^*) < \infty$, and

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\beta}\psi(Z_i;\beta) = \frac{1}{N}\sum_{i=1}^{N}D_i^*V_i^{*-1}W_iD_i^*.$$

Taking the second derivative,

$$\frac{\partial}{\partial\beta}(D_i^* V_i^{*-1} W_i D_i^*) = (\frac{\partial}{\partial\beta} D_i^{*\prime}) V_i^{*-1} W_i D_i^* + D_i^{*\prime}(\frac{\partial}{\partial\beta} V_i^{*-1}) W_i D_i^* + D_i^{*\prime} V_i^{*-1} W_i (\frac{\partial}{\partial\beta} D_i^*).$$

We also have

$$\frac{\partial}{\partial\beta} D_i^* = (1 - 2p_i) \otimes (\frac{\partial}{\partial\beta} A_i) X_i,$$

$$\frac{\partial}{\partial\beta} A_i = diag\left(\frac{X_{i1} e^{\beta' X_{i1}}(1 - e^{\beta' X_{i1}})}{(1 + e^{\beta' X_{i1}})^3}, ..., \frac{X_{iT} e^{\beta' X_{iT}}(1 - e^{\beta' X_{iT}})}{(1 + e^{\beta' X_{iT}})^3}\right).$$

Since $|X_i| < b < \infty$, $0 \leq (1 - 2p_i) \leq 1$, $0 \leq \frac{e^{\beta' X_{it}}}{(1+e^{\beta' X_{it}})^2} \leq \frac{1}{4}$, and $0 \leq \frac{1}{1+e^{\beta' X_{it}}} \leq 1$,

$$\frac{\partial}{\partial\beta} D_i^* < \infty.$$

And

$$\frac{\partial}{\partial\beta} V_i^{*-1} = -V_i^{*-1}(\frac{\partial}{\partial\beta} V_i^*) V_i^{*-1},$$

$$\frac{\partial}{\partial\beta} V_i^* = (\frac{\partial}{\partial\beta} A_i^{*\frac{1}{2}}) C_i(\gamma) A_i^{*\frac{1}{2}} + A_i^{*\frac{1}{2}} C_i(\gamma)(\frac{\partial}{\partial\beta} A_i^{*\frac{1}{2}}).$$

Then

$$\begin{aligned}\frac{\partial}{\partial\beta}A_i^{*\frac{1}{2}} &= diag[\frac{\partial}{\partial\beta}\sqrt{(1-2p_{i1})^2\frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2}+p_{i1}(1-p_{i1})},...,\\ &\quad \frac{\partial}{\partial\beta}\sqrt{(1-2p_{iT})^2\frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2}+p_{iT}(1-p_{iT})}]\\ &= \frac{1}{2}diag[\frac{(1-2p_{i1})^2\frac{X_{i1}e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2}-2\frac{X_{i1}e^{2\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^3}}{\sqrt{(1-2p_{i1})^2\frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2}+p_{i1}(1-p_{i1})}},...,\\ &\quad \frac{(1-2p_{iT})^2\frac{X_{iT}e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2}-2\frac{X_{iT}e^{2\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^3}}{\sqrt{(1-2p_{iT})^2\frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2}+p_{iT}(1-p_{iT})}}]\\ &= \frac{1}{2}diag[\frac{(1-2p_{i1})^2\mu_{i1}(1-\mu_{i1})(1-2\mu_{i1})X_{i1}}{\sqrt{(1-2p_{i1})^2\frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2}+p_{i1}(1-p_{i1})}},...,\\ &\quad \frac{(1-2p_{iT})^2\mu_{iT}(1-\mu_{iT})(1-2\mu_{iT})X_{iT}}{\sqrt{(1-2p_{iT})^2\frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2}+p_{iT}(1-p_{iT})}}].\end{aligned}$$

Therefore, $\frac{\partial}{\partial\beta}A_i^{*\frac{1}{2}} < \infty$, and $\frac{\partial}{\partial\beta}V_i^* < \infty$. Condition 6 is verified.

To verify condition 7 of Juan and Jennrich (1998) , the derivative of $\psi(Z_i;\beta)$ with respect to β is:

$$\begin{aligned}\frac{\partial}{\partial\beta}\psi(Z_i;\beta) &= \frac{\partial}{\partial\beta_t}(D_i^{*\prime}V_i^{*-1}W_i(Z_i-\mu_i^*))\\ &= D_i^*V_i^{*-1}W_iD_i^* + (\frac{\partial}{\partial\beta}D_i^*)'V_i^{*-1}W_i(Z_i-\mu_i^*) + D_i^{*\prime}(\frac{\partial}{\partial\beta}V_i^{*-1})W_i(Z_i-\mu_i^*).\end{aligned}$$

We already showed the following equations when we verifying condition 6,

$$E(\frac{\partial}{\partial\beta}\psi(Z_i;\beta)) = D_i^* V_i^{*-1} W_i D_i^*.$$

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\beta}\psi(Z_i;\beta) = \frac{1}{N}\sum_{i=1}^{N} D_i^* V_i^{*-1} W_i D_i^*.$$

To complete verifying condition 7, we need to show that

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\beta}\psi(Z_i;\beta) \to I_0^*(\beta).$$

almost surely as $N \to \infty$.

Since

$$\frac{1}{N}\sum_{i=1}^{N} D_i^* V_i^{*-1} D_i^* = \frac{1}{N}\sum_{i=1}^{N}((1-2p_i)\otimes A_i X_i')' V_i^{*-1} W_i (1-2p_i)\otimes A_i X_i',$$

and $(1-2p_i)$, V_i, A_i W_i are all bounded from previous proof. Then $((1-2p_i)\otimes A_i)' V_i^{*-1} W_i (1-2p_i)\otimes A_i$ is bounded below by a positive constant b_i.

Let a denote any $T \times 1$ vector, then

$$\frac{1}{N} a' \sum_{i=1}^{N} X_i((1-2p_i)\otimes A_i)' V_i^{*-1} W_i (1-2p_i)\otimes A_i X_i' a \geq \frac{1}{N} b_i a' \sum_{i=1}^{N} X_i X_i' > 0,$$

by Assumption C, which is

$$\frac{1}{N}\sum_{i=1}^{N} X_i X_i' \to B,$$

as $N \to \infty$, where B is a positive definite matrix.

Then,

$$\frac{\partial}{\partial\beta}\psi(Z_i;\beta) < \infty,$$

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial\beta}\psi(Z_i;\beta) \to I_0^*(\beta),$$

almost surely as $N \to \infty$.

To verify condition 8 of Juan and Jennrich (1998), we already show that each term of the second derivatives of $\psi(Z_i;\beta)$ with respect to β is bounded when we verify condition 6 ($\frac{\partial}{\partial\beta}D_i^* < \infty$, $\frac{\partial}{\partial\beta}V_i^{*-1} < \infty$).

Hence,

$$\frac{\partial^2}{\partial\beta\partial\beta}\psi(Z_i;\beta) < \infty.$$

In conclusion, condition 2 of Juan and Jennrich (1998) has been verified.

Liapounov's Theorem and Cramer-Wald Theorem are used to verify condition 3,

$$\sqrt{N}U_N\left(\beta_0\right) \underrightarrow{L} N\left(0, I_1^*(\beta_0)\right),$$

as $N \to \infty$.

As defined earlier,

$$U_N\left(\beta_0\right) = \frac{1}{N}\sum_{i=1}^{N}\psi(Z_i;\beta_0) = \frac{1}{N}\sum_{i=1}^{N} D_i^{*\prime}V_i^{*-1}W_i\left(Z_i - \mu_i^*\right).$$

Let a denote any $T \times 1$ vector, to apply Liapounov's Theorem, take

$$r_i = a' D_i^{*\prime}V_i^{*-1}W_i Z_i.$$

Then the mean of r_i is

$$m_i = E(r_i) = a' D_i^{*\prime}V_i^{*-1}W_i\mu_i^*,$$

and the variance of r_i is

$$Var(r_i) = a' D_i^{*\prime}V_i^{*-1}W_i A_i^* W_i V_i^{*-1} D_i^* a.$$

Define

$$c_n^2 = \sum_{i=1}^{N} Var(r_i) = \sum_{i=1}^{N} a' D_i^{*\prime}V_i^{*-1}W_i A_i^* W_i V_i^{*-1} D_i^* a = O(N),$$

since

$$\frac{1}{N}\sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} W_i A_i^* W_i V_i^{*-1} D_i^* \to I_1^*,$$

under condition 4.

Assume $E(|Z_i - \mu_i|^3) = \mu_{3i}^* < \infty$. Taking $\delta = 1$, the third central moment is:

$$\begin{aligned} E(|r_i - m_i|^3) &= E(|a' D_i^{*\prime} V_i^{*-1} W_i (Z_i - \mu_i^*)|^3) \\ &\leq (a' D_i^{*\prime} V_i^{*-1} W_i)^3 E(|Z_i - \mu_i^*|^3) \\ &= (a' D_i^{*\prime} V_i^{*-1} W_i)^3 \mu_{3i}^*. \end{aligned}$$

So

$$\sum_{i=1}^{N} E(|r_i - m_i|^3) = O(N),$$

since D_i^*, V_i^*, and W_i are bounded, which have been showed when verifying condition 4.

Then

$$\frac{1}{c_n^3}\sum_{i=1}^{N} E(|r_i - m_i|^3) = \frac{O(N)}{O(N^{\frac{3}{2}})} = O(N^{-1/2}),$$

which converges to zero as $N \to \infty$. Therefore, the conditions of Liapounov's theo-

rem are satisfied, and

$$\begin{aligned} T_N &= \frac{\sum_{i=1}^{N}(r_i - m_i)}{c_n} \\ &= \frac{\sum_{i=1}^{N} a' D_i^{*\prime} V_i^{*-1} W_i (Z_i - \mu_i^*)}{\sqrt{\sum_{i=1}^{N} a' D_i^{*\prime} V_i^{*-1} W_i A_i^* W_i V_i^{*-1} D_i^* a}} \\ &\underrightarrow{L} N(0,1), \end{aligned}$$

as $N \to \infty$.

By Slutsky's Theorem,

$$\sqrt{N} \sum_{i=1}^{N} a' D_i^{*\prime} V_i^{*-1} W_i (Z_i - \mu_i^*) \underrightarrow{L} N(0, a' I_1^*(\beta) a).$$

By the Cramer-Wold Theorem,

$$\sqrt{N} \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} W_i (Z_i - \mu_i^*) \underrightarrow{L} N(0, I_1^*(\beta)),$$

where $I_1^*(\beta) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (D_i^{*\prime} V_i^{*-1} W_i A_i^* W_i V_i^{*-1} D_i^*)$.

Thus, condition 3 has been verified.

3.8.2 Proof of consistency

Theorem 1. *Under Assumptions A-C, with probability one there exist zeros* $\hat{\beta}_N$ *of* $U_N(\beta) = 0$ *such that* $\hat{\beta}_N \to \beta_0$ *as* $N \to \infty$.

To prove **Theorem 1**, let $U_N(\beta)$ denote a sequence for which conditions 1 and 2 satisfied. By condition 1, $U_N(\beta_0) \to 0$ a.s., as $N \to \infty$. And by condition 2, $\frac{\partial}{\partial \beta^T} U_N(\beta)$ is nonsingular at β_0. Thus, β_0 is the unique zero of $U(\beta)$ in a neighborhood M of β_0.

Theorem 1 of Yuan and Jennrich (1998) states that under conditions 1 and 2, for any $\delta > 0$, there exists $\hat{\beta}_N \in M(\beta_0, \delta)$ such that $U_N\left(\hat{\beta}_N\right) = 0$ with probability 1, for all N sufficiently large.

By Theorem 1, there exists a zero $\hat{\beta}_N$ of $U_N(\beta)$ in $M(\beta_0, \delta)$ for all N sufficiently large. Let β^* be any limit point of $\hat{\beta}_N$, then $\beta^* \in M(\beta_0, \delta)$. Let $\hat{\beta}_{N_i}$ be any subsequence of $\hat{\beta}_N$, then $\hat{\beta}_{N_i} \to \beta^*$. Thus, $U_{N_i}\left(\hat{\beta}_{N_i}\right) \to U(\beta^*)$, and $U(\beta^*) = 0$.

Since β_0 is the only zero of $U(\beta)$ in a neighborhood $M(\beta_0, \delta)$, then $\beta^* = \beta_0$. Since this is true for all limit points of $\hat{\beta}_N$, $\hat{\beta}_N \to \beta_0$. Since conditions 1 and 2 hold with probability one, $\hat{\beta}_N \to \beta_0$ with probability one.

3.8.3 Proof of asymptotic normality

Theorem 2. *Under Assummptions A-C,* $\sqrt{N}\left(\hat{\beta}-\beta_0\right) \underset{L}{\Rightarrow} N\left(0, I_0^{*-1}(\beta_0) I_1^*(\beta_0) I_0^{*-1}(\beta_0)\right)$, *as* $N \to \infty$.

To prove **Theorem 2**, a Taylor series expansion of

$$\frac{1}{N} U_N\left(\hat{\beta}_N\right) = \frac{1}{N} \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} W_i \left(Z_i - \mu_i^*\right)$$

at β_0 yields

$$U_N\left(\hat{\beta}_N\right) = U_N\left(\beta_0\right) + \frac{\partial}{\partial \beta'} U_N\left(\beta_0\right)\left(\hat{\beta}_N - \beta_0\right) = 0.$$

Setting the expression above, and rearranging the terms, we get

$$\sqrt{N}\left(\hat{\beta}_N - \beta_0\right) \cong -\left(\frac{1}{N}\frac{\partial}{\partial \beta'} U_N\left(\beta_0\right)\right)^{-1} \frac{1}{\sqrt{N}} U_N\left(\beta_0\right).$$

To prove **Theorem 1**, we have already demonstrated that:

$$\frac{1}{N}\frac{\partial}{\partial \beta'} U_N\left(\beta_0\right) \to I_0^*(\beta_0),$$

almost surely as $N \to \infty$.

Since we already show that when verifying condition 3 of Yuan and Jennrich (1998),

$$\frac{1}{\sqrt{N}} U_N(\beta_0) \xrightarrow{L} N(0, I_1^*(\beta_0)).$$

as $N \to \infty$.

Then by Theorem 4 of Yuan and Jennrich (1998) and by slutsky's Theorem,

$$\sqrt{N}\left(\hat{\beta} - \beta_0\right) \xrightarrow{L} N\left(0, I_0^{*-1}(\beta_0) I_1^*(\beta_0) I_0^{*-1}(\beta_0)\right),$$

as $N \to \infty$.

And $\hat{I}_0^*$ and $\hat{I}_1^*$ can be estimated as:

$$\begin{aligned}\hat{I}_0^* &= \sum_{i=1}^{N} (\hat{D}_i^{*\prime} \hat{V}_i^{*-1} W_i \hat{D}_i^*) \\ &= \sum_{i=1}^{N} (1-2p_i) \otimes X_i' \hat{A}_i (\hat{A}_i^{*\frac{1}{2}} \hat{C}_i^*(\gamma) \hat{A}_i^{*\frac{1}{2}})^{-1} W_i (1-2p_i) \otimes \hat{A}_i X_i,\end{aligned}$$

$$\begin{aligned}\hat{I}_1^* &= \sum_{i=1}^{N} (\hat{D}_i^{*\prime} \hat{V}_i^{*-1} W_i \hat{A}_i^* W_i \hat{V}_i^{*-1} \hat{D}_i^*) \\ &= \sum_{i=1}^{N} (1-2p_i) \otimes X_i' \hat{A}_i (\hat{A}_i^{*\frac{1}{2}} \hat{C}_i^*(\gamma) \hat{A}_i^{*\frac{1}{2}})^{-1} W_i \hat{A}_i^* W_i (\hat{A}_i^{*\frac{1}{2}} \hat{C}_i^*(\gamma) \hat{A}_i^{*\frac{1}{2}})^{-1} (1-2p_i) \otimes \hat{A}_i X_i.\end{aligned}$$

3.9 References

Carroll KM. A Cognitive Behavioral Approach: Treating Cocaine Addiction. Rockville, MD: National Institute on Drug Abuse; 1998.

Fiellin D.A., Barry D.T., Sullivan L.E., Cutter C.J., Moore B.A., O'Connor P.G., Schottenfeld R.S. A randomized trial of cognitive behavioral therapy in primary care-based buprenorphine. Am J Med. 2013;126(1):74.

Fitzmaurice G.M., Molenberghs G., Lipsitz S.R. Regression models for longitudinal binary responses with informative Drop-outs. J.R.Statist. Soc. 1995; 57(4):691-704.

Fitzmaurice G.M., Davidian M, Verbeke G, Molenberghs G. Longitudinal Data Analysis. Chapman Hall/CRC: Boca Raton, FL, 2009.

Huang Y., Leroux B. Informative Cluster Sizes for Subcluster-Level Covariates and Weighted Generalized Estimating Equations. Biometrics. 2011; 67: 843-851.

Liang K., Zeger S. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73:13-22.

Lin C., Huiman X., Barnhart, Andrzej S., Kosinski. The Weighted Generalized Estimating Equations Approach for the Evaluation of Medical Diagnostic Test at Subunit Level. Biometrical Journal. 2006; 48(5): 758-771.

Lin H., Scharfstein D. O., Rosenheck R. A. Analysis of longitudinal data with irregular, outcome-dependent follow-up. Journal of the Royal Statistical Society. 2004; 66(3): 791-813.

Lipsitz S.R, Molenberghs G., Fitzmaurice G.M., Ibrahim J. GEE with Gaussian estimation of the correlations when data are incomplete. Biometrics. 2000; 56:528-536.

O'Hara Hines R.J., Hines W.G.S., Friesen T.G. A comparison of two drop-out weighting schemes in the analysis of clustered data with categorical and continuous responses. Journal of Agricultural, Biological, and Environmental Statistics.1999; 4(3):203-216.

Paik, M. C., Wang, C. Handling missing data by deleting completely observed records. J. Statist. Plann. Infer. 2009; 139:2341-2350.

Preisser J.S., Lohman K.K., Rathouz P.J. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. Statist. Med. 2002; 21: 3035-3054.

Robins, J. M., Rotnitzky, A., Zhao, P. Analysis of semiparametric regression models for re- peated outcomes in the presence of missing data. Journal of the American Statistical Association. 1995; 90:106-121.

Rotnitzky A., Wypij D. A note on the Bias of Estimators with missing data. Biometrics 1994; 50(4): 1163-1170.

Satty A., Mwambi H. Molenberghs G. Different methods for handling incomplete longitudinal binary outcome due to missing at random dropout. Statistical Methodology 2015; 24: 12-27.

Troxel A.B., Lipsitz S.R. Brennan T.A. Weighted Estimating Equations with Nonignorably Missing Response Data. Biometrics. 1997; 53(3):857-869.

Yuan K.H., Jennrich R.I. Asymptotics of Estimating Equations under Natural Conditions. Journal of Multivariate Analysis. 1998; 65: 245-260.

Chapter 4

Weighted Generalized Estimating Equations Approach for Longitudinal Binary Outcomes with Drop-outs Missing at Random

4.1 Introduction

Missing data are common in longitudinal studies. The cause of missing data may be due to subjects dropping out of the study or subjects returning to the study after being non-responsive for a while. A subject is called a drop-out when the re-

sponse variable is observed for certain visits and is missing for all consequent visits (Preisser et al. 2002). Previous research classified missing data mechanisms into three categories (Schafer and Graham. 2002). (1) Missing completely at random (MCAR)-where the probability of an observation missing does not depend on observed or unobserved measurements; (2) missing at random (MAR)-where given the observed data, the missing mechanism does not depend on the unobserved data; (3) missing not at random (MNAR)-where a missing observation depends on the unobserved data.

The Generalized Estimating Equations (GEE) estimators hold consistency if the data is MCAR, yet it can be subject to bias when the data is MAR, depending on the model's accuracies (Fitzmaurice et al. 1995). Additionally, GEE is often biased when applied to MNAR data, because the missing data is related to the unobserved responses and involves assumptions that cannot be tested within the data (Rotnitzky et al. 1998).

The Weighted Generalized Estimating Equations (WGEE) is an extension of the GEE approach, with a weight added to the GEE. WGEE has been widely used for analyzing incomplete longitudinal data, and gives consistent estimations under MAR when the dropout mechanism is correctly specified. The weight in the WGEE, which is estimated from some assumed dropout models, is usually defined as the inverse probability of being observed.

A recent study by Satty et al. (2015) compared the performance of three different methods for analyzing incomplete longitudinal binary outcome due to MAR; the methods were General Linear Mixed Models (GLMM), WGEE, and multiple imputation based on GEE (MIGEE). And they found that MIGEE performed better in both small and large sample sizes. Preisser et al. (2000) compared WGEE approach to a likelihood-based method to analyze the smoking trends with incomplete longitudinal binary response, the WGEE estimators perform better in large cluster sizes. Another study by Preisser et al. (2002) compared the performance of WGEE and GEE for longitudinal binary data with MAR drop-outs, and concluded that WGEE resulted in a smaller sample bias than GEE when the drop-out model was correctly specified. Chen et al. (2010) provided an approach to analyze longitudinal response and covariate data that are MAR using inverse probability WGEE. Lipsitz et al. conducted another simulation study for the analysis of a similar binary response dataset with missing, and concluded that the GEE model performed well under MCAR, while the consistency of GEE may not hold under MAR. On the other hand, the bias of the WGEE approach is negligible (Lipsitz et al. 2000). Other studies compared the WGEE estimators and the weighted least squares estimators under MAR assumption with simulation, and suggested that the WGEE outperformed the weighted least squares estimators, and remained consistent under various scenarios of missing data and sample sizes (Lin et al. 2006).

The purpose of this Chapter is to apply the inverse probability WGEE to a longitudinal dataset with MAR binary responses. The chapter is organized as follows. In

section 4.2, we present notation and model equations. In Section 4.3, we analyze the Self-reported Cocaine use and Urine test (SCU) data applying inverse probability WGEE approach. We provide the discussion and conclusion in Section 4.4.

4.2 Methods

From previous chapter, let Y_{it} denote the true drug use variable, and X_{it} be the covariate vectors for estimation at times $t = 1, ..., T$ for subjects $i = 1, ..., N$. Then, for the ith subject at time t, $Y_{it} = 1$ if the subject uses drug, $Y_{it} = 0$ if the subject does not use drug. Y_{it} is a binary response variable and its marginal distribution is Bernoulli:

$$f_y\left(y_i \mid X_i\right) = pr\left(Y_{i1} = y_1, ..., Y_{iT} = y_T \mid X_i\right) = exp(y_{it}\eta_{it} - log(1 + exp(\eta_{it}))).$$

The marginal mean of the drug use for the ith subject at a given time point t is denoted by μ_{it}. Let β be a vector of the regression parameters, then

$$\mu_{it} = E\left(Y_{it} \mid X_{it}, \beta\right) = Pr\left(Y_{it} = 1 \mid X_{it}, \beta\right),$$

logit link function will be used

$$\eta_{it} = \log \frac{\mu_{it}}{1 - \mu_{it}} = x_{it}\beta.$$

Liang and Zeger(1986) have proposed GEE of the form:

$$U_\beta(\beta) = \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}' V_{it}^{-1}(Y_{it} - \mu_{it}) = 0,$$

where $D_{it} = \partial\mu_{it}/\partial\beta$ and V_i is the covariance matrix of Y_i, which can be decomposed into the form $A_i^{\frac{1}{2}} C_i(\gamma) A_i^{\frac{1}{2}}$, where A_i is a matrix with the marginal variances on the main diagonal and zeros elsewhere, γ is a vector which fully characterize $C_i(\gamma)$, and $C_i(\gamma)$ is a working correlation matrix of Y_i's.

After forming μ_{it} to a vector $\mu_i = (\mu_{i1}, ...\mu_{iT})'$, we can write the GEE of the form:

$$U_\beta(\beta) = \sum_{i=1}^{N} D_i' V_i^{-1}(Y_i - \mu_i) = 0.$$

This estimate of β is consistent, if the data is MCAR, even if V_i, the covariance matrix of Y_i is misspecified. However, under MAR, the GEE approach may yield biased estimates.

On the other hand, the WGEE approach provides a consistent estimate of regression parameters under the assumption of MAR if the mean model and the missing mechanism of the model are correctly specified. Robins et al. (1995) has proposed WGEE of the form:

$$U_\beta(\beta) = \sum_{i=1}^{N} D_i' V_i^{-1} H_i(Y_i - \mu_i) = 0,$$

where $D_i = \partial\mu_i/\partial\beta$ and $V_i = A_i^{\frac{1}{2}} C_i(\gamma) A_i^{\frac{1}{2}}$ is the covariance matrix of Y_i, $H_i = diag(q_{i1}h_{i1}, ..., q_{it}h_{it})$ is the weighted matrix, where $q_{it} = 1$ if the outcome for subject

i is observed at time t; otherwise, $q_{it} = 0$. As a result, the weight H_{it} is h_{it} for an observed visit and 0 for an unobserved visit.

From the method proposed by Preisser et al.(2002), an inverse probability weight h_{it} can be obtained through a logistic regression model. The weight can later be used in the WGEE approach for parameter estimations. Under MAR missing mechanism, let Q_{it} be the indicator for observing the outcome at time t, and $\lambda_{it} = P(Q_{it} = 1|Q_{i(t-1)} = 1, X_{it}, Y_{it}, \theta)$ be the probability of observing the outcome at time t for the ith individual conditional on the individual being observed at the previous time point $t-1$. For the first time point, assume $Q_{i1} = 1$ and $\lambda_{i1} = 1$.

$\hat{\lambda}_{it}$ can be estimated by fitting a logistic model, $logit(\lambda_{it}(\theta)) = T_{it}\theta$, with a vector of predictors, T_{it}, which may include indicator variables of visit, covariates, and past response variables. After taking differentiation with respect to θ of the log partial likelihood, we have the score equation of θ for the ith subject as:

$$S_i(\theta) = \sum_{t=1}^{T} Q_{i,t-1} T_{it}(Q_{it} - \lambda_{it}(\theta)) = 0.$$

By solving the above score equation, we can obtain $\hat{\theta}$ and $\hat{\lambda}_{it}$.

The weight h_{it} is then defined as the inverse of the unconditional probability of being observed at time t, which can be estimated by the conditional probability,

$$\hat{h}_{it} = \left(\hat{\lambda}_{i1} \times ... \times \hat{\lambda}_{it}\right)^{-1}.$$

In this approach, an observation with a low probability of being observed will receive a large weight.

$\hat{\beta}$ is estimated through an iterative algorithm from a modified Fisher scoring for β by solving the WGEE. Given an initial guess of $\hat{\beta}^0$, update $\hat{\beta}^p$ in the pth iteration by taking:

$$\hat{\beta}^{p+1} = \hat{\beta}^p - \left(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{H}_i \hat{D}_i \right)^{-1} \left(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{H}_i \left(Y_i - \hat{\mu}_i \right) \right).$$

Under correctly specified models for the marginal means and for the missing mechanisms, WGEE provides a consistent estimate of β, which has an asymptotic normal distribution. The form of the sandwich estimator for asymptotic variance of $\hat{\beta}$ is:

$$\left(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{H}_i \hat{D}_i \right)^{-1} \left(\sum_{i=1}^{N} \hat{G}_i \hat{G}_i' \right) \left(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{H}_i \hat{D}_i \right)^{-1},$$

where $\hat{G}_i = \hat{U}_i - (\sum_{i=1}^{N} \hat{U}_i \hat{S}_i')(\sum_{i=1}^{N} \hat{S}_i \hat{S}_i') \hat{S}_i$, $\hat{U}_i = \hat{D}_i' \hat{V}_i^{-1} \hat{H}_i (Y_i - \hat{\mu}_i)$ is the weighted equation, and $\hat{S}_i = \sum_{t=1}^{T} Q_{i,t-1} T_{it} (Q_{it} - \hat{\lambda}_{it}(\hat{\theta}))$ is the score equation of $\hat{\theta}$ (Robins et al. 1995).

As stated in previous chapters, the self-reported data Z_{it} may be contaminated and deviated from the true data Y_{it}. The equation of Z_{it} is:

$$Z_{it} = Y_{it}\left(1 - R_{it}\right) + \left(1 - Y_{it}\right) R_{it},$$

where R_{it} represents an indicate variable for outcome contamination at times $t = 1, ..., T$, for subjects $i = 1, ..., N$, indicating whether self-reported data is the same as true drug use or not. $R_{it} = 1$ if there exists contamination, i.e. self reported data is not the same as true drug use data, otherwise $R_{it} = 0$.

Following the methods proposed in Chapter 2, the MCGEE form of the self-reported data Z_i can be written as:

$$U_\beta^*\left(\beta\right) = \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1}\left(Z_i - \mu_i^*\right) = 0,$$

where $\mu_i^* = E(Z_i|X_i, \beta) = (\mu_{i1}^*, ...\mu_{iT}^*)'$ is the expected value of Z_i, and the expected value of indicate variable for contamination R_i is denoted by p_i,

$$\begin{aligned}\mu_{it}^* = E(Z_{it}|X_{it}, \beta) &= E(Y_{it}|X_{it}, \beta) \times E((1 - R_{it})) + E((1 - Y_{it})) \times E(R_{it}) \\ &= \mu_{it} - 2\mu_{it} \times p_{it} + p_{it}.\end{aligned}$$

And,

$$\begin{aligned} D_i^* &= \partial\mu_i^*/\partial\beta = (1-2p_i) \otimes \frac{\partial\mu_i}{\partial\beta} \\ &= (1-2p_i) \otimes \frac{e^{x_i\beta}}{(1+e^{x_i\beta})^2} X_i \\ &= (1-2p_i) \otimes A_i X_i, \end{aligned}$$

where $A_i = diag(var(Y_{i1}), ..., var(Y_{iT}))$, $var(Y_{it}) = \mu_{it} \times (1-\mu_{it}) = \frac{e^{x_{it}\beta}}{(1+e^{x_{it}\beta})^2}$. $\otimes$ means that only multiply the tth row of vector $1-2p_i$ by the same tth row of matrix A_iX_i, i.e., $(1-2p_i) \otimes A_iX_i = ((1-2p_{i1}) \times a_{i1}x_{i1}, ..., (1-2p_{iT}) \times a_{iT}x_{iT})'$. V_i^* is the covariance matrix of Z_i, which can be decomposed into the form $A_i^{*\frac{1}{2}} C_i^*(\gamma) A_i^{*\frac{1}{2}}$, where A_i^* is a matrix with the marginal variances on the main diagonal and zeros elsewhere, i.e., $A_i^* = diag(var(Z_{i1}), ..., var(Z_{iT}))$, and

$$\begin{aligned} var(Z_{it}) &= \mu_{it}^*(1-\mu_{it}^*) \\ &= (\mu_{it} - 2\mu_{it}p_{it} + p_{it})(1 - \mu_{it} + 2\mu_{it}p_{it} - p_{it}) \\ &= \mu_{it} - 2\mu_{it}p_{it} + p_{it} - \mu_{it}^2 + 2\mu_{it}^2 p_{it} - p_{it}\mu_{it} + 2\mu_{it}^2 p_{it} - 4\mu_{it}^2 p_{it}^2 + 2\mu_{it}p_{it}^2 - \mu_{it}p_{it} \\ &\quad + 2\mu_{it}p_{it}^2 - p_{it}^2 \\ &= (1-2p_{it})^2 \mu_{it}(1-\mu_{it}) + p_{it}(1-p_{it}) \\ &= (1-2p_{it})^2 var(Y_{it}) + p_{it}(1-p_{it}). \end{aligned}$$

γ is a vector which fully characterize $C_i^*(\gamma)$, and $C_i^*(\gamma)$ is a working correlation matrix of Z_i.

After adding the inverse probability of drop-out as the weight for missing data, we can write the model equation as:

$$U_{\beta}^{*}(\beta) = \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} H_i (Z_i - \mu_i^{*}) = 0.$$

where $H_i = diag(q_{i1}h_{i1}, ..., q_{it}h_{it})$ is the weighted matrix, and $q_{it} = 1$ if the outcome for subject i is observed at time t; otherwise, $q_{it} = 0$.

We also proposed a Mean Corrected WGEE approach to correct the report bias of Z_{it} in Chapter 3, which adds an inverse probability of contamination as the weight into the MCGEE. To count the missing data under this approach, the model equation is:

$$U_{\beta}^{*}(\beta) = \sum_{i=1}^{N} D_i^{*\prime} V_i^{*-1} H_i^{*} (Z_i - \mu_i^{*}) = 0,$$

where $H_i^{*} = diag(q_{i1}h_{i1}^{*}, ..., q_{it}h_{it}^{*})$ is the weighted matrix, and $q_{it} = 1$ if the outcome for subject i is observed at time t; otherwise, $q_{it} = 0$. h_{it}^{*} can be estimated as:

$$\hat{h}_{it}^{*} = \left(\hat{\lambda}_{i1} \times ... \times \hat{\lambda}_{it}\right)^{-1} \times (\hat{p}_{it})^{-1},$$

where λ_{it} is the probability of observing the outcome at time t for the ith individual conditional on the individual being observed at the previous time point $t-1$, and p_{it} is the contamination probability for subject i at time t. p_{it} can be estimated using the two methods proposed in Chapter 3.

4.3 Results

4.3.1 Data description

In the Self-reported Cocaine use and Urine test (SCU) data, there are a total of 140 patients, followed for a period of 5-6 months. After a 2-week induction and stabilization period, during which patients were treated by nurses 3 times per week with 16 mg buprenorphine daily, enrolled subjects were randomly assigned to the treatment or the control group. Both groups received buprenorphine, a substitute for cocaine use, which was stored in bottles. Buprenorphine was instructed to be taken once per day. If the bottle was opened on a specific day, the patient was regarded as adherent. The special MEMSCAP bottles can record the time when the bottle is opened.

The control group received physical management (PM), a 15-20 minutes session by Internal Medicine physicians with experiences as buprenorphine providers. Throughout the study period, sessions occurred weekly for the first two weeks, every two weeks for the next four weeks and then monthly. The treatment group received PM plus cognitive behavioral therapy (CBT). CBT is a counseling intervention that has demonstrated efficacy in treating a variety of psychiatric conditions and cocaine dependences. CBT was provided by masters- and doctoral-level clinicians who were trained with a manual adapted from a guidance for the use of CBT for cocaine dependence (Carroll 1998). The main components of counseling focused on performing a functional analysis of behavior, promoting behavioral activation, identifying and coping with drug cravings, enhancing drug refusal skills and decision makings about

high risk situations, and improving problem solving skills (Fiellin et al. 2013).

The study's major outcomes include (1) self-reported daily cocaine uses which were reported during the weekly PM sessions, and (2) weekly urine cocaine test results. Our main interest is to test the difference of the self-reported cocaine use between treatment group and control group.

Some patients may be transferred for protective purposes. One of the criteria for transfer is three consecutive weeks of positive urine tests for drug use (missing urine screenings are counted as positive) after the buprenorphine dose increased to 24mg daily. Patients may also be transferred for psychiatric or other medical problems, as well as other continued drug use, such as Benzos. The transfer was a clinical decision made by the patients' primary clinicians in the study (Fiellin et al. 2013).

Dropout is another main issue for participants. There are various reasons for dropouts: the patients may realize that the counseling isn't working; they may feel guilty if they are continuing to use; they may not like the therapist; or they may not like talking about their problems. Among the 69 patients in the treatment group, 34 of them dropped out or transfered from the study, while 36 out of 71 patients in the control group dropped out or transfered (Table 4.1). We assume missing data from the transfers and the dropouts is MAR.

Analyses are performed using the R software.

Table 4.1: Number of Completed and Dropout Patients for each group

	Treatment	Control	Total
Completed	35	35	70
Dropouts or transfers	34	36	70
Total	69	71	140

4.3.2 Results

Table 4.2 shows the differences between the characteristics of present and absent subjects at the end of the study. There are no observed differences of drop-out between the treatment and control group. For the percentage of self-reported cocaine use days within the follow-up periods for each subject, the mean value is 10.9% for people who are absent and 9.4% for people who are not. Subjects who dropped out tend to have reported cocaine use on more days, but this difference is not considered significant (p-value=0.2925). The mean percentage of the Buprenorphine bottle open days during the study periods for each absent subject at the end of the study is 52.3% and for present subjects at the end of the study is 68.9%. Because this difference is significant (p-value=0.0001), we can conclude that adherent patients are less likely to drop out, suggesting that Buprenorphine bottle open data can be used for missing prediction.

Prior to fitting the models, we first calculate the weights to be used in the WGEE

Table 4.2: Comparison among characteristics of the study subjects present and absent at the end of the study

			End of the Study		
		Total $N = 140$	Present $N = 70$	Absent $N = 70$	p-value
Group	Treatment	69(49.3%)	35(50%)	34(48.6%)	1
	Control	71(50.7%)	35(50%)	36(51.4%)	
Cocaine use days%	Mean (SD)	10.1%(8%)	9.4%(8%)	10.9%(9%)	0.2925
Bottle open days%	Mean (SD)	60.6%(26%)	68.9%(25%)	52.3%(25%)	0.0001*

approach by implementing a logistic regression model for the missing indicators. The predictor of missing data only includes the percentage of Buprenorphine bottle open days during the study period for each subject, since it has significant effect on drop-outs. The model equation is:

$$\log \frac{Pr\left(\lambda_i = 1 | T_i, \theta\right)}{1 - Pr\left(\lambda_i = 1 | T_i, \theta\right)} = \theta_0 + \theta_1 \times bottleopendays\%.$$

Table 4.3: Results of missing indicator analysis

	Estimate	S.E.	P-values
θ_0	-1.60	0.49	0.001
θ_1	2.62	0.73	0.0004

The missing weight can be estimated as the inverse of the conditional probability of being observed, $\hat{h}_i = \left(\hat{\lambda}_i\right)^{-1}$. The results are presented in Table 4.3.

Based on the approaches discussed in Chapter 2 and Chapter 3, we add these weights to the MCGEE approach and the MCWGEE approach under several assumptions. We conduct our analysis using these models with the logistic link:

$$\log \frac{Pr\left(Z_{it}=1|X_i,\beta\right)}{1-Pr\left(Z_{it}=1|X_i,\beta\right)}=\beta_0+\beta_1 X_i.$$

First, we assume the contamination probability does not depend on the MEMSCAP bottle open data and is estimated using the first approach described in section 2.2.2. After adding a missing weight to the MCGEE approach, the results are presented in Table 4.4 and 4.5.

Table 4.4: Results of MCGEE of cocaine use under the MAR assumption (h=1)

	Estimate	S.E.	P-value
β_0	-2.72	0.13	<0.0001
β_1	0.39	0.18	0.03

Table 4.5: Results of MCGEE of cocaine use under the MAR assumption (h=4)

	Estimate	S.E.	P-value
β_0	-2.43	0.17	<0.0001
β_1	0.31	0.20	0.12

After adding the inverse probability of being observed to the subject specific MCWGEE approach under the same assumption, results are shown in the following tables.

Table 4.6: Results of MCWGEE of cocaine use under the MAR assumption (h=1)

	Estimate	S.E.	P-value
β_0	-2.97	0.14	<0.0001
β_1	0.33	0.19	0.08

Table 4.7: Results of MCWGEE of cocaine use under MAR assumption of missing (h=4)

	Estimate	S.E.	P-value
β_0	-2.40	0.17	<0.0001
β_1	0.37	0.18	0.04

The estimates of the CBT effect seem similar for MCGEE and MCWGEE under the assumption that contamination indicator doesn't depend on the bottle open data for both $h = 1$ and $h = 4$ cases. And these estimates increase slightly when compared to the results in Chapter 2 and Chapter 3.

Second, we assume the contamination indicator depends on the MEMSCAP bottle open data, and then add the missing weight to the MCGEE and MCWGEE approach under $h = 1$ and $h = 4$.

Table 4.8: Results of MCGEE of cocaine use under the MAR assumption (h=1)

	Estimate	S.E.	P-value
β_0	-2.21	0.12	<0.0001
β_1	0.19	0.15	0.20

Table 4.9: Results of MCGEE of cocaine use under the MAR assumption (h=4)

	Estimate	S.E.	P-value
β_0	-2.29	0.11	<0.0001
β_1	0.19	0.15	0.20

From Table 4.8 - 4.11, we observe that the estimators of the effect of CBT for all

Table 4.10: Results of MCWGEE of cocaine use under the MAR assumption (h=1)

	Estimate	S.E.	P-value
β_0	-2.18	0.11	<0.0001
β_1	0.20	0.15	0.18

Table 4.11: Results of MCWGEE of cocaine use under the MAR assumption (h=4)

	Estimate	S.E.	P-value
β_0	-2.24	0.11	<0.0001
β_1	0.20	0.15	0.18

these methods are very similar, however, when compared with the results in Chapter 2 and 3, the estimators change more significantly after adding the inverse probability of being observed as the weight.

4.4 Discussion and Conclusion

Previous studies have proposed WGEE approach with weight as the estimated probability of dropout at the time of attrition (Robins et al. 1995). This approach yielded consistent estimates when the responses were MAR and the probability of dropout had been correctly specified (Fitzmaurice et al. 1995). We applied this methods together with the mean corrected estimating equations to the SCU data to investigate the impacts of patient's protective transfers and dropouts.

The impact of MAR longitudinal binary outcome depends on the frequency of miss-

ing data, and the association between the missing indicators and the binary response variables (Chen et al. 2010). 50% of the patients dropped out or transferred at the end of the study. We have shown that the percentage of Buprenorphine bottle open days during the study periods has significant effect on drop-outs among several missing data indicators. We build a logistic regression model of this significant missing indicator to estimate the conditional probability of being observed at the end of the study, and estimate the missing weight as the inverse of this conditional probability. Finally, we add the estimated missing weight to the MCGEE and MCWGEE approach we have proposed in the previous chapters.

After applying this approach on the SCU data to address the issue of dropouts, we find that the effects of CBT have all changed when compared to the previous models under different circumstances. Under the assumption that contamination indicator doesn't depend on the bottle open data, these estimates seem to increase after adding the missing weight. When the time period for cocaine to be cleared from urine is 1 day ($h = 1$), the estimate of CBT effect (β_1) increases from 0.29 to 0.39 for MCGEE approach, and increases from 0.31 to 0.33 for MCWGEE approach. After the time period for cocaine to be cleared from urine is increased to 4 days ($h = 4$), the estimate of CBT effect (β_1) increases from 0.30 to 0.31 for MCGEE approach, and increases from 0.35 to 0.37 for MCWGEE approach.

However, under the assumption that contamination indicator depends on the bottle open data [the contamination probability estimation is built on a model which in-

cludes the MEMSCAP bottle open data], these estimates decrease after adding the missing weight. Under the assumption that the time period for cocaine to be cleared from urine is 1 day ($h = 1$), the estimate of CBT effect (β_1) decreases from 0.27 to 0.19 for MCGEE approach, and decreases from 0.28 to 0.20 for MCWGEE approach. Under the assumption that the time period for cocaine to be cleared from urine is 4 days ($h = 4$), the estimate of CBT effect (β_1) decreases from 0.27 to 0.19 for MCGEE approach, and decreases from 0.28 to 0.20 for MCWGEE approach. In our analysis, assuming both the missing weight and the contamination depend on the bottle open data resulted in more significant changes of the estimates of the treatment effect.

As a summary of the SCU data analysis, the results of MCGEE and MCWGEE approach without missing weight are similar. After including the missing weight, we find that for the first simple assumption, the estimation for contamination probability and the missing weight are basically independent, since we primarily used bottle open data to estimate the drop-out probability, and did not include this information in detecting the contamination in the self-reported data. For the second model assumption, both the estimation for contamination probability and the missing weight depend mainly on the bottle open data. This could explain the significant differences in the estimators.

4.5 References

Carroll KM. A Cognitive Behavioral Approach: Treating Cocaine Addiction. Rockville, MD: National Institute on Drug Abuse; 1998.

Chen B., Yi G.Y., Cook R.J. Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. Journal of the American Statistical Association. 2010; 105(489): 336-353.

Fiellin D.A., Barry D.T., Sullivan L.E., Cutter C.J., Moore B.A., O'Connor P.G., Schottenfeld R.S. A randomized trial of cognitive behavioral therapy in primary care-based buprenorphine. Am J Med. 2013;126(1):74.

Fitzmaurice G.M., Molenberghs G., Lipsitz S.R. Regression models for longitudinal binary responses with informative Drop-outs. J.R.Statist. Soc. 1995; 57(4):691-704.

Liang K., Zeger S. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73:13-22.

Lin C., Huiman X., Barnhart, Andrzej S., Kosinski. The Weighted Generalized Estimating Equations Approach for the Evaluation of Medical Diagnostic Test at Subunit Level. Biometrical Journal. 2006; 48(5): 758-771.

Lipsitz S.R, Molenberghs G., Fitzmaurice G.M., Ibrahim J. GEE with Gaussian es-

timation of the correlations when data are incomplete. Biometrics. 2000; 56:528-536.

Preisser J.S., Galecki A.T., Lohman K.K., Wagenknecht L.E. Analysis of Smoking Trends with Incomplete Longitudinal Binary Responses. Journal of the American Statistical Association. 2000; 95(452): 1021-1031.

Preisser J.S., Lohman K.K., Rathouz P.J. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. Statist. Med. 2002; 21: 3035-3054.

Robins, J. M., Rotnitzky, A., Zhao, P. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association. 1995; 90:106-121.

Rotnitzky A., Wypij D. A note on the Bias of Estimators with missing data. Biometrics 1994; 50(4): 1163-1170.

Rotnitzky A., Robins J.M., Scharfstein D.O. Semiparametric regression for repeated outcomes with nonignorable nonresponse. Journal of the American Statistical Association 1998;93:1321-1339.

Satty A., Mwambi H., Molenberghs G. Different methods for handling incomplete longitudinal binary outcome due to missing at random dropout. Statistical Method-

ology 2015; 24: 12-27.

Schafer J., Graham J. Missing data: our view of the state of the art. Psychological Methods. 2002; 7: 147-177.