CONDITIONAL WEIGHTED GENERALIZED QUASILIKELIHOOD INFERENCES IN INCOMPLETE LONGITUDINAL MODELS FOR BINARY AND COUNT DATA

TASLIM S. MALLICK

Conditional Weighted Generalized Quasilikelihood Inferences in Incomplete Longitudinal Models for Binary and Count Data

by

©Taslim S. Mallick

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirement for the Degree of Doctor of Philosophy in Statistics

Department of Mathematics and Statistics
Memorial University of Newfoundland

April, 2009

St. John's Newfoundland Canada

Abstract

There exists an inverse probability weight (INPW) based unconditional estimating equation approach (a correction to accommodate the missingness nature of the data) for computing unbiased regression estimates in an incomplete longitudinal set-up mainly for binary data. It is however known that this INPW based unconditional estimating equation approach still may produce regression estimates with large bias. It is demonstrated in this thesis that it would be much better to use an INPW based conditional estimating equation approach to obtain unbiased and hence consistent estimates for the regression effects. This approach however requires the longitudinal correlation structure to be known. Under the assumption that the binary or count data follow an autoregressive order-1 [AR(1)] type model, the thesis develops a conditional weighted generalized quasilikelihood (CWGQL) approach that accommodates both missingness and the longitudinal correlation issues properly. This appears to be a major improvement over the existing INPW based generalized estimating equation (GEE) approach which either fails to use the longitudinal correlations or uses 'working' correlations approach. Extensive simulation studies are undertaken to examine the relative performance of the proposed CWGQL approach with the existing INPW based GEE approach. Finally the incomplete longitudinal models are generalized to study the survey based incomplete longitudinal data. A stratified finite population is considered to examine the performance of a stratified random sampling (StRS) based CWGQL approach in estimating the regression parameters involved in the finite population for both binary and count data models.

Acknowledgements

A few lines are too short to make a complete account of my deep appreciation for my supervisor Professor B.C. Sutradhar. With his vast knowledge and deep insight in the subject, he guided and inspired me throughout the course of my Ph.D. program. Without his inspirational guidance, his enthusiasm, his encouragements, his unselfish help, I could never finish this work. I owe to him what I know in this area.

I am grateful to my co-supervisor Dr. Gary Sneddon for his guidance, continued encouragement and invaluable suggestions. It has been a distinct privilege for me to work with both Drs. Sutradhar and Sneddon for which I shall forever be grateful.

I sincerely acknowledge the financial support provided by the School of Graduate Studies, Department of Mathematics and Statistics and my supervisors in the form of Graduate Fellowships and Teaching Assistantships. Further, I wish to thank the Department for providing us a friendly atmosphere and the necessary facilities to complete my program.

I am also thankful to my supervisor for arranging an internship program for four months with Statistics Canada where I gained valuable insight to deal with complex data structure under the guidence of Dr. Milorad Kovacevic. This internship program was supported by MITACS, NPCDS and Statistics Canada. My sincere thanks are due to them.

I would also like to thank the examiners of the thesis: Drs. Richard J. Cook, Alwell Oyet and Zhao Zhi Fan for their comments and suggestions.

I am grateful to my parents, my wife, brother and sister for their eternal love, emotional support and encouragement during this program.

It is my great pleasure to thank my friends and well-wishers who directly or indirectly encouraged and helped me during my Ph.D. program.

Contents

List of Tables

Chapter 1

Introduction

1.1 Motivation of the Problem

It is well known that there is no unique way to model the correlations of longitudinal discrete data such as binary and count, even if the data are complete, that is, the data do not contain any missing responses. This has drawn renewed interest among the researchers for the analysis of such complete longitudinal discrete data. Statistical inference gets more complicated when one considers more practical issues such as the possible incompleteness in the longitudinal binary or count data. Note that there also exist work over the last two decades dealing with longitudinal missing data. But, these studies do not appear to address both correlation and missingness issues properly. As far as the incompleteness is concerned, it is mostly expected that the data would be (a) missing completely at random (MCAR) or (b) missing at random (MAR) [Rubin (1976), Paik (1997), Fitzmaurice, Laird and Zahner (1996)]. The inference for the data MAR is however complicated. We will review the definitions of these incompleteness and their existing uses in Chapter 2.

We now in brief, provide a review of some of the widely used recent models for correlated binary and count data and also review the related inference issues for such complete longitudinal data.

Suppose that in a longitudinal study there are K independent individuals. Also suppose that had there been no missing data, $y_i = (y_{i1}, \cdots, y_{it}, \cdots, y_{iT})'$ and $x_i = (x_{i1}, \cdots, x_{it}, \cdots, x_{iT})'$ with $x_{it} = (x_{it1}, \cdots, x_{itu}, \cdots, x_{itp})'$ denote the $T \times 1$ complete outcome (either binary or count) vector and $T \times p$ covariate matrix, respectively, recorded from the i-th ($i = 1, \cdots, K$) individual over T successive points in time. Since the repeated responses of the same individual are likely to be correlated, any inferences about the regression effects without accommodating the correlation structure for the repeated responses would lead to inefficient estimates. But, as mentioned earlier, there is no unique way to model correlations for repeated binary or count responses.

In the following two sections, we briefly review the existing correlation models for the repeated binary and count responses and point out their advantages and drawbacks. Note that our objective would be choosing a suitable model among them such that the selected model can be as general as possible accommodating one or more correlation structures, with a wide range for the possible correlations.

1.1.1 A review of some existing correlated binary models

For known x_{it}, let $\mu_{it} = E(Y_{it} \mid x_{it}) = exp(x_{it}'\beta)/[1+exp(x_{it}'\beta)]$, where $\beta = (\beta_1, \cdots, \beta_p)'$ is the p-dimensional regression effects of y_{it} on x_{it}. For convenience, we will however use $E(Y_{it})$ for $E(Y_{it} \mid x_{it})$ throughout the thesis. Also let $\rho_{i|t-t'|}$ be the $|t - t'|$-th lag correlation between the two binary responses y_{it} and $y_{it'}$. That is $corr(y_{it}, y_{it'}) = \rho_{itt'} \equiv \rho_{i|t-t'|}$. We use these notations and write the following correlated binary models.

(a) Bahadur model

Bahadur (1961) has proposed the multidimensional binary distribution given by

$$f(y_{i1}, \cdots, y_{iT} \mid x_{i1}, \ldots, x_{iT}) \quad = \quad \prod_{t=1}^{T} \mu_{it}^{y_{it}} (\bar{\mu}_{it})^{1-y_{it}} \left[1 + \sum_{t<t'}^{T} \rho_{i|t-t'|} (\mu_{it}\bar{\mu}_{it}\mu_{it'}\bar{\mu}_{it'})^{-1/2} \right.$$

$$\times(y_{it}-\mu_{it})(y_{it'}-\mu_{it'})\Bigg], \tag{1.1}$$

where $\bar{\mu}_{it} = 1 - \mu_{it}$. After some algebra one may show that this model (1.1) yields

$$E(Y_t) = \mu_{it},\ V(Y_{it}) = \mu_{it}\bar{\mu}_{it},\ \text{and}\, corr(Y_{it}, Y_{it'}) = \rho_{i|t-t'|}. \tag{1.2}$$

Note that for simplicity, we will use $f(y_{i1}, \cdots, y_{iT})$ for $f(y_{i1}, \cdots, y_{iT} \mid x_{i1}, \ldots, x_{iT})$ throughout the thesis. Further note that the lag correlations $\rho_{i|t-t'|}$ used in the probability model (1.1) may accommodate a class of Gaussian type auto-correlation structures. For example, for autoregressive order-1 [AR(1)] type correlations, one uses $\rho_{i|t-t'|} = \rho_i^{|t-t'|}$ and for moving average order-1 [MA(1)], $\rho_{i1} = \rho_i$ and $\rho_{i|t-t'|} = 0$ if $|t-t'| \neq 1$. This property that one may use a suitable correlation structure in (1.1) is certainly an advantage for this model. But, the range of correlations provided by this model depends on the marginal probabilities. For example, for $T = 2$, ρ_{i1} satisfies

$$max\left[-\left(\frac{\mu_{i1}\mu_{i2}}{\bar{\mu}_{i1}\bar{\mu}_{i2}}\right)^{1/2}, -\left(\frac{\bar{\mu}_{i1}\bar{\mu}_{i2}}{\mu_{i1}\mu_{i2}}\right)^{1/2}\right] < \rho_{i1} < min\left[\left(\frac{\mu_{i1}\bar{\mu}_{i2}}{\mu_{i2}\bar{\mu}_{i2}}\right)^{1/2}, \left(\frac{\mu_{i2}\bar{\mu}_{i2}}{\mu_{i1}\bar{\mu}_{i2}}\right)^{1/2}\right].$$

In general

$$max_{y_{i1},\cdots,y_{iT}}\left[-\frac{1}{f_1^*(y_{i1},\cdots,y_{iT})}\right] < \rho_{i|t-t'|}$$

$$< min_{y_{i1},\cdots,y_{iT}}\left[\frac{1-f_1^*(y_{i1},\cdots,y_{iT})}{f_1^*(y_{i1},\cdots,y_{iT})f_2^*(y_{i1},\cdots,y_{iT})}\right], \tag{1.3}$$

where $f_1^*(y_{i1},\cdots,y_{iT}) = \prod_{t=1}^T \mu_{it}^{y_{it}}(\bar{\mu}_{it})^{1-y_{it}}$, $f_2^*(y_{i1},\cdots,y_{iT}) = \sum_{t<t'}^T (y_{it}-\mu_{it})(y_{it'}-\mu_{it'})/[\mu_{it}\bar{\mu}_{it}\mu_{it'}\bar{\mu}_{it'}]^{1/2}$ and $max_{y_{i1},\cdots,y_{iT}}[g(.)]$ or $min_{y_{i1},\cdots,y_{iT}}[g(.)]$ denote the maximum or minimum of the function $g(.)$ for all possible values of $y_{i1},\cdots,y_{iT}$.

Note that these range restrictions shown above limit the use of this correlation model in practice, even though some authors such as Cox (1972) and Prentice (1988) have applied this model for certain data analysis. To have some more specific idea about the range restriction implied by the Bahadur model (1.1), we refer to Farrell and Sutradhar (2006). For example, Farrell and Sutradhar (2006) have shown that for the stationary case with $T = 4$ and $\rho_{i|t-t'|} = \rho_i^{|t-t'|}$, $\rho_{i1} = \rho_i$ lies in the range $-0.262 < \rho_i < 0.449$ when $\mu_{it} = \mu_i = 0.4$. Note that this range is quite narrow. In general the ranges produced by the Bahadur model (1.1) are quite narrow. This limits the use of the Bahadur model (1.1) in practice.

(b) Kanter model

Kanter (1975) introduced an observation-driven correlation model for stationary binary data. Sutradhar (2008) has extended Kanter's model to the non-stationary longitudinal mixed model set-up. For the longitudinal non-stationary binary data, Kanter's model can be written as

$$y_{it} = s_{it}\{y_{i,t-1} \oplus d_{it}\} + (1 - s_{it})d_{it}, \tag{1.4}$$

where s_{it} follows a binary distribution with

$$P(s_{it} = 1) = \eta \tag{1.5}$$

and d_{it} follows another binary distribution with

$$P(d_{it} = 1) = \frac{\mu_{it} - \eta\mu_{i,t-1}}{1 - 2\eta\mu_{i,t-1}} = \xi_{it}. \tag{1.6}$$

In (1.4), $\oplus$ denotes addition modulo 2. For binary $y_{i,t-1}$ with $P(y_{i,t-1} = 1) = \mu_{i,t-1}$ and assuming that $y_{i,t-1}$, s_{it} and d_{it} are independent, it can be shown that y_{it} follows the binary distribution with $E(Y_{it}) = \mu_{it}$ and $V(Y_{it}) = \mu_{it}\bar{\mu}_{it}$. The computation of the auto-covariances between two binary responses, however, depend on the value of T. For example, when $T = 4$

$$\begin{aligned} cov(Y_{i2}, Y_{i4}) &= E(Y_{i2}Y_{i4}) - \mu_{i2}\mu_{i4} \\ &= P(Y_{i2} = 1, Y_{i4} = 1) - \mu_{i2}\mu_{i4}, \end{aligned}$$

where the computation of the joint $P(Y_{i2} = 1, Y_{i4} = 1)$ may be obtained as

$$P(Y_{i2} = 1, Y_{i4} = 1) = P(Y_{i2} = 1)\,[P(Y_{i3} = 1 \mid Y_{i2} = 1)P(Y_{i4} = 1 \mid Y_{i3} = 1)$$

$$
\begin{aligned}
& +P(Y_{i3}=0 \mid Y_{i2}=1)P(Y_{i4}=1 \mid Y_{i3}=0)] \\
= \;& \mu_{i2}\left[\{\eta+(1-2\eta)\xi_{i3}\}\{\eta+(1-2\eta)\xi_{i4}\}\right. \\
& \left.+\{(1-\eta)-(1-2\eta)\xi_{i3}\}\xi_{i4}\right]. \qquad (1.7)
\end{aligned}
$$

In general, the auto-correlation can be computed by

$$\rho_{i|t-t'|} = \frac{cov(Y_{it}, Y_{it'})}{\sqrt{V(Y_{it})V(Y_{it'})}}.$$

Note that for the stationary case, when $\mu_{it} = \mu_i$, ξ_{it} reduces to

$$\xi_i = P(d_{it} = 1) = \frac{\mu_i(1-\eta)}{1-2\eta\mu_i},$$

where η is given in (1.5). For this special case it may be shown that $\rho_{i|t-t'|} = \rho_i^{|t-t'|}$, with $\rho_i = \eta(1-2\mu_i)/(1-2\eta\mu_i)$. Further note that since $0 < \xi_i < 1$, ρ_i must satisfy the range restriction through η as $0 < \eta < min[(1-\mu_i)/\mu_i, 1]$. For example, for $T = 4$ and $\mu_{it} = \mu_i = 0.4$, it follows that $0 < \rho_i < 1$ [Farrell and Sutradhar (2006)]. Note that this range is wider than that of the Bahadur model (1.1) which makes the Kanter model (1.4) some what relaxed for its practical use. Note however that Kanter's model (1.4) is an AR(1) type model. The MA(1) or equicorrelation (EQC) type models will have different forms than that of (1.4). Consequently, Kanter's model (1.4) is not so general as the Bahadur model (1.1) which can accommodate a class of correlation structures.

(c) A conditional linear dynamic model

Qaqish (2003) has used a family of multivariate binary distributions through a linear dynamic conditional probability given that the marginal means and correlations are specified. This conditional linear family is given by

$$
\begin{aligned}
\lambda_{it|t-1,\cdots,1}(y^*_{i,t-1}) &= P(Y_{it} = 1 \mid y^*_{i,t-1}) = E(Y_{it} \mid y^*_{i,t-1}) \\
&= \mu_{it} + \sum_{j=1}^{t-1} b_{itj}(y_{ij} - \mu_{ij}), \qquad (1.8)
\end{aligned}
$$

where $y^*_{i,t-1} = (y_{i1}, \cdots, y_{i,t-1})'$ and $b^*_{i,t-1} = (b_{it1}, \cdots, b_{it,t-1})'$ is computed based on the specified correlation structure or using

$$b^*_i = [cov(Y^*_{i,t-1})]^{-1} cov(Y^*_{i,t-1}, Y_{it}).$$

The use of the correlation structure becomes clear when $cov(Y^*_{i,t-1})$ is expressed as $cov(Y^*_{i,t-1}) = A_i^{1/2} C_i A_i^{1/2}$, where $A_i = diag(a_{i11}, \cdots, a_{i,t-1,t-1})$ with $a_{itt} = V(Y_{it})$, and $C_i = (c_{itt'})$ is a suitable correlation structure. Note that similar to the Bahadur model, one may use Gaussian type AR(1), MA(1) and EQC correlation structure to define this C_i matrix. This is a practical advantage of the conditional linear model (1.8).

Note, however, that since $0 < \lambda_{it|t-1,\cdots,1}(y^*_{i,t-1}) < 1$ in (1.8), the ranges for the correlations $(c_{itt'})$ are bound to be restricted. For example, if the correlation matrix C_i takes the AR(1) form, namely, $C_i = (c_{itt'}) = \rho_i^{|t-t'|}$ for all $t \neq t'$, then the correlation parameter ρ_i is bounded as $\rho_i \geq max(-\psi_i^2, -1/\psi_i^2)$, where $\psi_i = \sqrt{\mu_i/(1-\mu_i)}$, $\mu_i = \mu_{i1} = \cdots = \mu_{iT}$ being the stationary mean. In such a case with $T = 4$ and $\mu_i = 0.4$, ρ_i satisfies the restriction $-0.667 < \rho_i < 1$ [Farrell and Sutradhar (2006)]. Note that this range is clearly the widest when compared with the ranges for the similar correlation parameter under the Bahadur and Kanter's models.

(d) A conditional non-linear dynamic model

Recall that all three correlated binary models discussed above yield the same marginal mean $\mu_{it} = P(Y_{it} = 1) = exp(x'_{it}\beta)/[1 + exp(x'_{it}\beta)]$ and the same marginal variance $\mu_{it}(1 - \mu_{it})$. If we are, however, willing to consider a binary model with a dynamic mean structure such as the mean at a given time being a function of the past means, we may then use a non-linear binary dynamic model which yields the lag correlations with full range, i.e., $-1 < \rho_{i|t-t'|} < 1$. For convenience, we write one of this type of non-linear dynamic models as follows

$$P(Y_{it} = 1 \mid y_{i,t-1}) = \frac{exp(x'_{it}\beta + \eta_1 y_{i,t-1})}{1 + exp(x'_{it}\beta + \eta_1 y_{i,t-1})}, \tag{1.9}$$

where η_1 is the dynamic dependence parameter. Note that this model has been extensively used in the past in the econometrics literature [Amemiya (1985), Manski (1987)] and most recently has been used by Sutradhar and Farrell (2007) [see also Farrell and Sutradhar (2006)], among others.

In the present thesis, our objective is to examine the effects of longitudinal correlation structure as well as non-responses on the inferences about the regression effects involved in the correlated binary and Poisson models with fixed marginal means. As far as the longitudinal correlation structure is concerned, we will use (c), the conditional linear binary dynamic (CLBD) model discussed above. This is because this CLBD model produces fixed marginal means as opposed to the dynamic (recursive) marginal means produced by the non-linear model (d). The CLBD model also produces correlations with widest possible ranges as compared to the other two linear models, namely, the Bahadur and Kanter's models.

1.1.2 A review of the existing correlated Poisson models

Unlike for the correlated binary models, not much attention has been given to modelling the correlated Poisson data, especially when the count data are collected repeatedly over time. Some of the early works, see for example, Johnson and Kotz (1969, Chapter 11, Section 4) [see also Holgate (1964), Campbell (1934), Teicher (1954), and Dwass and Teicher(1957)] dealt with a specialized correlated Poisson model, where the clustered or repeated counts are assumed to follow an EQC structure. To understand this structure, let y_{i0} follow the Poisson distribution with mean parameter μ^*_{i0}. That is

$$f(y_{i0}) = \frac{e^{-\mu^*_{i0}} {\mu^*_{i0}}^{y_{i0}}}{y_{i0}!} \tag{1.10}$$

which we denote for convenience by $y_{i0} \sim P(\mu^*_{i0})$. Also let $y^*_{it} \sim P(\mu^*_{it})$ and $y^*_{i0}, y^*_{i1}, \cdots, y^*_{iT}$ are independent. Now suppose that the repeated counts $y_{i1}, \cdots, y_{iT}$ are generated following

$$y_{it} = y^*_{it} + y^*_{i0}, \text{ for } t = 1, \cdots, T. \tag{1.11}$$

It is clear from (1.11) that y_{it} $(t = 1, \cdots, T)$ marginally follows the Poisson distribution with parameter $(\mu^*_{i0} + \mu^*_{it})$, and jointly they are correlated with pairwise lag correlations

$$\begin{aligned} corr(Y_{it}, Y_{it'}) &= \frac{cov(Y_{it}, Y_{it'})}{\sqrt{V(Y_{it})V(Y_{it'})}} \\ &= \frac{\mu^*_{i0}}{\sqrt{(\mu^*_{it} + \mu^*_{i0})(\mu^*_{it'} + \mu^*_{i0})}}. \end{aligned} \tag{1.12}$$

Note that if the data are stationary, that is, $\mu^*_{i0} = \mu^*_{i1} = \cdots = \mu^*_{iT} = \mu_i$, then y_{it} and $y_{it'}$ have the constant correlation 1/2. Thus, this correlation model (1.11)-(1.12) is heavily restricted when the data are stationary.

Further note that whether the data are stationary or non-stationary, the correlation model (1.11) produces a complicated joint distribution for the repeated responses $y_{i1}, \cdots, y_{iT}$. This may be understood easily from the bivariate case with $T = 2$, where the probability distribution of y_{i1} and y_{i2} has the form

$$p(y_{i1}, y_{i2}) = e^{-(\mu^*_{i0}+\mu^*_{i1}+\mu^*_{i2})} \sum_{j=0}^{min(y_{i1}, y_{i2})} \frac{\mu^{*\,j}_{i0}}{j!} \frac{\mu^{*\,y_{i1}-j}_{i1}}{(y_{i1}-j)!} \frac{\mu^{*\,y_{i2}-j}_{i2}}{(y_{i2}-j)!} \tag{1.13}$$

[Johnson and Kotz (1969, eq. 52, p.298)], which is complicated for the likelihood inference purpose for any parameters involved in μ^*_{i1} and μ^*_{i2}.

Note that the Poisson EQC model (1.11) has limited use in the longitudinal setup. This is because, even if the data are stationary, one would expect a variable time effect causing non-constant correlations, whereas under the model (1.11), all lag correlations are constant and equal to 1/2.

Recently, Sutradhar (2003) [see also McKenzie (1988)] has proposed a class of auto-correlations for the stationary longitudinal count data. The correlation models under such a class produce Gaussian type lag correlations. In the following subsection, we review these correlation models in brief.

1.1.2.1 A general stationary auto-correlation model for repeated count data

(a) A stationary Gaussian AR(1) type model

Let the responses y_{it} at time t be related to $y_{i,t-1}$ at time $t-1$ as

$$y_{it} = \rho * y_{i,t-1} + d_{it},\ t = 2, \ldots, T \tag{1.14}$$

[Sutradhar (2003), McKenzie (1988)], where for the given count $y_{i,t-1}$, $\rho * y_{i,t-1}$ denotes the binomial thinning operation defined as

$$\rho * y_{i,t-1} = \sum_{j=1}^{y_{i,t-1}} b_j(\rho), \tag{1.15}$$

with $P[b_j(\rho) = 1] = \rho$ and $P[b_j(\rho) = 0] = 1 - \rho$. We now assume that the covariates are time independent, that is, $x_{it} = x_i$ for all $t = 1, \cdots, T$, where x_{it} is the covariate value at time t for the i-th individual. We further assume that

(1) $y_{i1} \sim P(\mu_i)$ with $\mu_i = exp(x_i^{'}\beta)$

(2) $d_{it} \sim P[\mu_i(1-\rho)]$, and

(3) $y_{i,t-1}$ and d_{it} are independent

One may then show that

$$E(Y_{it}) = V(Y_{it}) = \mu_{it} \text{ and}$$

$$corr(Y_{it}, Y_{it'}) = \rho^{|t-t'|}. \tag{1.16}$$

Note that the correlations in (1.16) have similar structure as that of the Gaussian AR(1) model, and the correlation parameter ρ has the range $0 \leq \rho \leq 1$. Here the correlations exhibit a decaying pattern. To be specific, the correlations produced by model (1.14) appear to decay exponentially when the lag increases, whereas the correlations yielded by the model (1.11) are constant and equal to 1/2.

(b) A stationary Gaussian MA(1) type model

Under this model y_{it} is represented as the function of the present and past d_{it} as

$$y_{it} = \rho * d_{i,t-1} + d_{it}, \tag{1.17}$$

[Sutradhar (2003)]. Here, similar to (1.15), $\rho * d_{i,t-1} = \sum_{j=1}^{d_{i,t-1}} b_j(\rho)$. Now assuming $d_{it} \overset{iid}{\sim} P[\mu_i/(1+\rho)]$ for all $t = 1, \cdots, T$, it may be shown that the model (1.17) produces marginal means and variances as μ_i, and produces Gaussian MA(1) type correlation structure. To be specific, the correlation structure under such a model is given by

$$corr(Y_{it}, Y_{it'}) = \begin{cases} \rho/(1+\rho), & \text{for } |t-t'| = 1 \\ 0, & \text{otherwise} \end{cases}, \tag{1.18}$$

which is similar to that of Gaussian type MA(1) structure, except unlike the Gaussian case where lag 1 correlation ranges between -1/2 to 1/2, the present restriction produces lag 1 correlation between 0 to 1/2.

(c) A stationary Gaussian EQC type model

An EQC type model for the count data may be expressed in the fashion similar to those of AR(1) type model (a) and MA(1) type model (b). To be specific, the model is written as

$$y_{it} = \rho * y_{i0} + d_{it}, \tag{1.19}$$

[Sutradhar (2003)]. Assuming $y_{i0} \sim P(\mu_i)$ and $d_{it} \overset{iid}{\sim} P(\mu_i(1-\rho))$ for all $t = 1, \cdots, T$, one may show that y_{it} in (1.19) marginally follows a Poisson distribution with stationary mean parameter μ_i. Moreover, it can be shown that

$$corr(Y_{it}, Y_{it'}) = \rho, \tag{1.20}$$

for all $t \neq t'$. Note that for this Gaussian type EQC structure produced by (1.19), ρ however lies in the range $0 \leq \rho \leq 1$ instead of $-1/(T-1) \leq \rho \leq 1$ under the Gaussian EQC model. Further note that the Poisson additive model (1.11) produces the same EQC structure as (1.20) only when one assumes $\mu_{i0}^* = \mu_i\rho \neq \mu_{it}^* = \mu_i(1-\rho)$ for $t = 1, \cdots, T$.

1.1.2.2 Some remarks on non-stationary correlation models

Note that in practice, the covariates may be time dependent causing non-stationarity in the means and the variances. To construct such non-stationary models as a generalization of the stationary models (1.14), (1.17) and (1.19), we refer to Sutradhar, Jowaheer and Sneddon (2008), for example, and explain one of the non-stationary models [AR(1)] as follows:

In order to generalize the stationary Poisson AR(1) model (1.14) to the non-stationary case, first assume that the covariates x_{it} are time dependent. The relationship between y_{it} and $y_{i,t-1}$ written for the stationary case remains the same under the non-stationary case. Nevertheless, the new model produces non-stationary correlations as opposed to the stationary model. To be specific, the non-stationary Poisson AR(1) model has the same form

$$y_{it} = \rho * y_{i,t-1} + d_{it}, \text{ for } t = 2, \cdots, T \tag{1.21}$$

[Sutradhar et al. (2008)] as that of the stationary model (1.14), but the assumptions for those two models are different. The model (1.21) becomes non-stationary under the assumptions that

(i) $y_{i1} \sim P(\mu_{i1})$ with $\mu_{it} = exp(x_{it}^{'}\beta)$,

(ii) $d_{it} \sim P(\mu_{it} - \rho\mu_{i,t-1})$,

(iii) $y_{i,t-1}$ and d_{it} are independent.

One may then show that

$$E(Y_t) = V(Y_{it}) = \mu_{it} = exp(x_{it}^{'}\beta) \text{ and}$$

$$corr(Y_{it}, Y_{it'}) = \rho^{(t'-t)}\left[\frac{\mu_{it}}{\mu_{it'}}\right]^{1/2} \text{ for } t < t'. \tag{1.22}$$

Note that the marginal means and the variances and the lag $(t-t')$ correlations given in (1.22) are non-stationary in nature. This is because all of them are a function of μ_{it} which depends on the time dependent covariate x_{it}. Further note that the correlation structure in (1.22) reduces to Gaussian AR(1) type auto-correlation structure (1.16) under the stationary Poisson model (1.14) with $\mu_{it} = \mu_i = exp(x_i'\beta)$. As far as the range of the correlation parameter ρ is concerned, it is clear under the non-stationary case that $0 < \rho < min_{(i,t)}[1, \mu_{it}/\mu_{i,t-1}]$. This is obvious from the fact that the mean parameter $(\mu_{it} - \rho\mu_{i,t-1}$) of the Poisson random variable d_{it} must be positive.

In the present thesis, we will assume that the longitudinal correlation structures for both binary and count data are known. To be specific, we will consider AR(1) non-stationary models for both binary and count data. There are several reasons for such a consideration. First, it is most likely in practice that the lag correlations get smaller as the lag increases. This situation is accommodated by the non-stationary AR(1) models. Secondly, there does not appear to be any adequate discussion on the effects of longitudinal correlations on the inferences for longitudinally missing data. The use of a specified correlation model such as AR(1) non-stationary model is expected to reveal clear understanding about such inferences for the longitudinally missing data. If, the correlation structure is however unknown but belongs to a class as discussed above, one may follow Sutradhar et al. (2008) and study the missingness effects accordingly, which is however beyond the scope of the present thesis.

1.1.3 Complete data based estimation of parameters in non-stationary AR(1) models

Irrespective of whether the complete data is binary or count, one may estimate the regression paramter β by using the generalized quasilikelihood (GQL) approach proposed by Sutradhar (2003). Note that this GQL approach exploits mean, variance and covariances to estimate β, which is considered to be a generalization of the popular QL approach of Wedderburn (1974) [see also McCullagh (1983)], where the estimation is carried out exploiting the means and variances for the independent data. Further note that Sutradhar (2003) considered GQL estimation of β for stationary longitudinal data under a class of auto-correlation structures, whereas, as mentioned above, in the present section, we have chosen the non-stationary AR(1) correlation models only.

1.1.3.1 GQL estimation of the regression effects β

Suppose that the repeated responses, whether binary or count, have a non-stationary correlation structure defined as

$$C_i = (c_{itt'}) : T \times T$$

with $c_{itt'}$ as a known suitable function of ρ, x_{it} and $x_{it'}$, which, for convenience, we express as

$$c_{itt'} = h(\rho, x_{it}, x_{it'}), \tag{1.23}$$

'h' being a known suitable function. To be specific, for the non-stationary Poisson AR(1) model (1.21), $c_{itt'}$ has the form given by (1.22). For a non-stationary binary correlation model (1.27) to be discussed in the next subsection, the form of $c_{itt'}$ is shown in (1.31). Further, let $A_i = diag[\sigma_{i11}, \cdots, \sigma_{itt}, \cdots, \sigma_{iTT}]$ with $\sigma_{itt} = V(Y_{it})$. Under the Poisson AR(1) model (1.21) $\sigma_{itt} = \mu_{it} = exp(x'_{it}\beta)$, and under the AR(1) type binary model (1.27) σ_{itt} has the formula $\sigma_{itt} = V(Y_{it}) = \mu_{it}(1 - \mu_{it})$ with $\mu_{it} = exp(x'_{it}\beta)/[1 + exp(x'_{it}\beta)]$.

Now, for $\Sigma_i = A_i^{1/2} C_i A_i^{1/2}$, one may write the GQL estimating equation for β as

$$\sum_{i=1}^{K} \frac{\partial \mu'_i}{\partial \beta} \Sigma_i^{-1}(y_i - \mu_i) = \sum_{i-1}^{K} X'_i A_i \Sigma_i^{-1}(y_i - \mu_i) = 0, \tag{1.24}$$

[Sutradhar (2003)] where $X'_i = (x_{i1}, \cdots, x_{it}, \cdots, x_{iT})$ is the $(p \times T)$ covariate matrix, with $x_{it} = (x_{it1}, \cdots, x_{itu}, \cdots, x_{itp})'$. Here (1.24) is an unbiased estimating equation, because of the fact that for $X'_i A_i \Sigma_i^{-1} = B_i = (b_{iut})$ (say), $E(b_{iu_1} Y_{i1} + \cdots + b_{iu_T} Y_{iT}) = b_{iu_1}\mu_{i1} + \cdots + b_{iu_T}\mu_{iT}$ for all $u = 1, \ldots, p$, where $\mu_{it} = E(Y_{it})$.

Note that in (1.24), $y_i = (y_{i1}, \cdots, y_{iT})'$ and $\mu_i = (\mu_{i1}, \cdots, \mu_{iT})'$ as mentioned earlier. Further note that when the ρ parameter in C_i matrix involved in the GQL estimating equation (1.24) is consistently estimated (say, by a method of moments), the GQL estimate obtained from (1.24) becomes highly efficient, the maximum likelihood estimator (MLE) being fully efficient which is however extremely complicated to compute especially under the longitudinal count model (1.21).

1.1.3.2 Non-stationary AR(1) models and estimation of correlation index parameter ρ

(a) Correlated count data model

Note that under the longitudinal count data model (1.21), the non-stationary correlations $c_{itt'}$ are given by

$$c_{itt'} = \rho^{(t'-t)} \left[\frac{\mu_{it}}{\mu_{it'}}\right]^{1/2} \quad \text{for } t < t'$$

[as shown in (1.22)]. Let $\tilde{y}_{it} = (y_{it} - \mu_{it})/\sqrt{\sigma_{itt}}$. The correlation parameter ρ in (1.22) may be estimated by the method of moments (MM). The lag 1 response based MM estimator of the ρ parameter is given by

$$\hat{\rho} = \frac{\sum_{i=1}^{K}\sum_{t=2}^{T-1} \tilde{y}_{it}\tilde{y}_{i,t-1}}{\sum_{i=1}^{K}\sum_{t=1}^{T} \tilde{y}_{it}^2} \frac{KT}{\sum_{i=1}^{K}\sum_{t=2}^{T-1} [\sigma_{i,t-1,t-1}/\sigma_{i,tt}]^{1/2}}. \tag{1.25}$$

[See Mallick and Sutradhar (2008, eqn. 35) for similar moment estimation in the context of AR(1) type time series for count data]. In (1.25), $\sigma_{itt} = V(Y_{it}) = \mu_{it} = exp(x'_{it}\beta)$. Note that under the present AR(1) model for the count data, the correlation parameter has to satisfy the range $0 < \rho < min_{(i,t)}[1, \mu_{it}/\mu_{i,t-1}]$.

(b) Linear dynamic binary correlation model

Here, we consider a special case of the CLBD model (1.8) discussed in Section 1.1.1. We write this specialized non-stationary AR(1) model as

$$P(Y_{i1} = 1) \;\; = \;\; \mu_{i1} = exp(x'_{i1}\beta)/[1 + exp(x'_{i1}\beta)],$$

$$\begin{aligned} P(Y_{it} = 1 \mid y_{i,t-1}) &= \lambda_{i,t|t-1}(y_{i,t-1}) \\ &= \mu_{it} + \rho(y_{i,t-1} - \mu_{i,t-1}), \text{ for } t = 2, \cdots, T. \end{aligned} \tag{1.26}$$

Since μ_{it} and $\mu_{i,t-1}$ are the functions of β, for convenience we rewrite the conditional probability function $\lambda_{i,t|t-1}(y_{i,t-1})$ as

$$\lambda_{i,t|t-1}(\beta, \rho) = \mu_{it} + \rho(y_{i,t-1} - \mu_{i,t-1}). \tag{1.27}$$

It follows from (1.27) that

$$\begin{aligned} E(Y_{i2}) &= E_{Y_{i1}} E(Y_{i2} \mid Y_{i1}) \\ &= E_{Y_{i1}} [\mu_{i2} + \rho(Y_{i1} - \mu_{i1})] \\ &= \mu_{i2}, \end{aligned}$$

by the property that $E_{Y_{i1}}(Y_{i1}) = \mu_{i1}$. It then implies recursively that

$$E(Y_{it}) = \mu_{it} = exp(x'_{it}\beta)/[1 + exp(x'_{it}\beta)]. \tag{1.28}$$

Similarly

$$\begin{aligned} var(Y_{it}) &= E(Y_{it}^2) - [E(Y_{it})]^2 \\ &= E(Y_{it}) - [E(Y_{it})]^2 \\ &= \mu_{it}(1 - \mu_{it}) = \sigma_{i,tt}, \end{aligned} \tag{1.29}$$

by (1.28) for all $t = 1, \cdots, T$. Next, the covariances may be obtained in the similar fashion. For example,

$$\begin{aligned} cov(Y_{it}, Y_{i,t-1}) &= E(Y_{it}Y_{i,t-1}) - \mu_{it}\mu_{i,t-1} \\ &= E_{Y_{i,t-1}} [Y_{i,t-1} E(Y_{it} \mid Y_{i,t-1})] - \mu_{it}\mu_{i,t-1} \\ &= E_{Y_{i,t-1}} [Y_{i,t-1}\{\mu_{it} + \rho(y_{i,t-1} - \mu_{i,t-1})\}] \\ &= \rho\mu_{i,t-1}(1 - \mu_{i,t-1}) = \sigma_{i,t-1,t-1}. \end{aligned} \tag{1.30}$$

It is easy to verify that under this special model (1.27), the correlations between Y_{it} and $Y_{it'}$ for all $t, t' = 1, \cdots, T$ are given by

$$corr(Y_{it}, Y_{it'}) = c_{itt'} = \begin{cases} \rho^{t'-t} \left[\frac{\sigma_{itt}}{\sigma_{it't'}}\right]^{1/2}, & \text{for } t < t' \\ \rho^{t-t'} \left[\frac{\sigma_{it't'}}{\sigma_{itt}}\right]^{1/2}, & \text{for } t > t' \end{cases}. \tag{1.31}$$

Note that the mean, variance and correlation produced by the model (1.27) are non-stationary as all of them are functions of time dependent covariates $\{x_{it}\}$. Further note that the formula for the moment estimator of the correlation parameter ρ remains the same as given by (1.25) under the Poisson AR(1) model with $\sigma_{itt} = \mu_{it} = exp(x'_{it}\beta)$, whereas $\sigma_{itt} = \mu_{it}(1 - \mu_{it})$ with $\mu_{it} = exp(x'_{it}\beta)/[1 + exp(x'_{it}\beta)]$ under the present non-stationary binary AR(1) model (1.27). Furthermore, the correlation parameter ρ under the present non-stationary binary AR(1) model has to satisfy the range restriction

$$max\left[-\frac{\mu_{it}}{1-\mu_{i,t-1}}, -\frac{1-\mu_{it}}{\mu_{i,t-1}}\right] \leq \rho \leq min\left[\frac{1-\mu_{it}}{1-\mu_{i,t-1}}, \frac{\mu_{it}}{\mu_{i,t-1}}\right], \tag{1.32}$$

whereas under the count data model (1.21), ρ satisfies the range restriction $0 < \rho < min_{(i,t)}[1, \mu_{it}/\mu_{i,t-1}]$.

1.2 Objective of the Thesis

Note that in a longitudinal data analysis, it may happen in practice that a few responses are missing for some of the individuals under study. More specifically, the responses may be MCAR (missing completely at random) or most likely MAR (missing at random). If the longitudinal data are MAR but the inferences about the regression effects are made by treating the longitudinal responses as complete (or MCAR) as discussed in Section 1.1.3.1, it causes a serious biasness in the regression estimates. This inferential problem has attracted many researchers over the last two decades and some progress has been made toward obtaining unbiased regression effects by accommodating the missing mechanism in the inference procedure. A careful review of the existing literature however shows that most of the studies attempted to develop certain estimating equations for β which still may not be unbiased to produce unbiased estimates for β. Furthermore, the longitudinal correlation models are also not incorporated properly in developing such estimating equations.

The main objective of the thesis is to develop a new conditional unbiased estimating equation approach that accommodates both MAR mechanism as well as longitudinal correlations into account. The plan of the thesis is as follows

In Chapter 2, we discuss the advantages and drawbacks of the existing estimation approaches for the regression effects involved in incomplete longitudinal binary data models. In the same chapter, we also provide the proposed conditional weighted generalized quasilikelihood (CWGQL) approach for the estimation of the regression effects. Furthermore, an extensive simulation study is conducted to examine the relative performances of some of the existing estimation approaches as compared to the proposed approach.

In Chapter 3, we consider the inferences in an incomplete longitudinal count data set-up. Similar to that of the analysis for the incomplete binary data, we provide a clear comparison between the existing estimating equations and the proposed CWGQL estimating equations approaches. More specifically, it is demonstrated clearly how one can accommodate the longitudinal correlation structure for the count data in such incomplete count data analysis. We then conduct a simulation study to examine the small sample properties of the proposed as well as existing estimators.

In Chapter 4, as compared to Chapters 2 and 3, we consider a slightly more complex incomplete longitudinal data set-up, where it is assumed that the independent subjects are no longer selected randomly, instead, a complex survey design is used for their selection in the study. To reflect this sampling design for the collection of the longitudinal data, we modify the proposed CWGQL approach to accommodate this additional design issue. We then examine the performance of a design based CWGQL (DBCWGQL) approach in estimating the regression effects. Note that in Chapter 4, we concentrate only on the design based incomplete longitudinal binary data set-up.

In Chapter 5, we carry out the DBCWGQL inferences as in Chapter 4, but deal with design based incomplete longitudinal count data. This thesis concludes in Chapter 6.

Chapter 2

Incomplete Longitudinal Binary Model

In some of the longitudinal studies, it may happen that a few responses from some individuals are missing during the data collection period. Let R_{it} be a response indicator variable at time t $(t = 1, \cdots, T)$ for the i-th $(i = 1, \cdots, K)$ individual, so that

$$R_{it} = \begin{cases} 1, & \text{if } y_{it} \text{ is observed} \\ 0, & \text{otherwise.} \end{cases} \quad (2.1)$$

Note that it is quite appropriate to assume that all individuals in the longitudinal study provide the responses at the first time point $t = 1$. Thus, in notation,

$$R_{i1} = 1 \text{ for all } i = 1, \cdots, K.$$

Under the assumption that had there been no missing response, the longitudinal binary data would follow the probability model given in (1.26). Thus, the first response y_{i1} for the i-th individual follows a binary distribution with parameter

$$\mu_{i1} = P(Y_{i1} = 1) = exp(x'_{i1}\beta)/[1 + exp(x'_{i1}\beta)], \text{ denoted by } y_{i1} \sim bin(\mu_{i1}),$$

for all $i = 1, \cdots, K$.

2.1 Missing Data Process Beyond the First Response

Note that the longitudinal responses for $t = 2, \cdots, T$ can be missing either in an intermittent fashion or monotonically. For simplicity, we however assume in the thesis that the missingness occur in a monotonic pattern only. That is, the response indicators satisfy the following relationship

$$R_{i1} \geq R_{i2} \geq \cdots \geq R_{it} \geq \cdots \geq R_{iT}.$$

Let $y_i^c = (y_{i1}, \cdots, y_{iT})'$ now denote the complete data vector for the i-th individual and x_i^c is the corresponding complete covariate matrix. We however assume that x_i^c is known even if some of the responses are missing. Thus we will use x_i for x_i^c. Given $R_i = (R_{i1}, \cdots, R_{iT})'$, the complete data vector y_i^c can be partitioned as $y_i^c = (y_{oi}, y_{mi})$, where y_{oi} are the values of y_i^c that are observed and y_{mi} denotes the components of y_i^c that are missing. Next, let $\alpha = (\alpha_1, \cdots, \alpha_q)$ denote the vector of parameters of the non-response model so that $P(R_i \mid y_i^c, x_i, \alpha)$ denote the probability distribution of R_i given y_i^c and α. Here $x_i = (x_{i1}, \cdots, x_{it}, \cdots, x_{iT})'$ is the $T \times p$ covariate matrix with x_{it} is the p-dimensional covariate vector corresponding to y_{it}. In this notation, the responses are MCAR if

$$P(R_i \mid y_i^c, x_i, \alpha) = P(R_i \mid x_i, \alpha) \tag{2.2}$$

(i.e., missingness does not depend on the values of the data y_i^c) and they are MAR if

$$P(R_i \mid y_i^c, x_i, \alpha) = P(R_i \mid y_{oi}, x_i, \alpha) \tag{2.3}$$

(i.e., missingness depends only on the components y_{oi} of y_i^c that are observed, and not on the component that are missing). Finally, the missing data mechanism is nonignorable, if

$$P(R_i \mid y_i^c, x_i, \alpha) = P(R_i \mid y_{oi}, y_{mi}, x_i, \alpha) \tag{2.4}$$

that is, the probability of non-response depends on the missing values, y_{mi}, and/or unobserved responses. In the monotonic missing response case, one may illustrate the above three models by

$$M1: \quad P(R_{it} = 1 \mid y_i^c, x_i, R_{i,t-1} = 1) = P(R_{it} = 1 \mid x_i, R_{i,t-1} = 1),$$

$$M2: \quad P(R_{it} = 1 \mid y_i^c, x_i, R_{i,t-1} = 1) = P(R_{it} = 1 \mid y_{i1}, \cdots, y_{i,t-1}, x_i, R_{i,t-1} = 1),$$

$$M3: \quad P(R_{it} = 1 \mid y_i^c, x_i, R_{i,t-1} = 1) = P(R_{it} = 1 \mid y_{i1}, \cdots, y_{it}, \cdots, y_{iT}, x_i, R_{i,t-1} = 1),$$

respectively [Fitzmaurice et al. (1996), Paik (1997), Rubin (1976)].

Note that it is known that the inferences based on MCAR mechanism remains the same as those of complete data analysis. This is because, under this MCAR mechanism the response indicators have nothing to do with the structure of the responses, implying that the MCAR based data is simply a subset of the complete data with different sample size. The problem arises when data are MAR or nonignorable. For practical importance as well as for simplicity, we however deal with the MAR case (2.3) only. Furthermore, for clarity, we write the conditional distribution of R_{it} as

$$P(R_{it} \mid y_{oi}, x_i, \alpha) = \Big[P(R_{it} = 1 \mid H_{i,t-1})\Big]^{R_{it}} \Big[1 - P(R_{it} = 1 \mid H_{i,t-1})\Big]^{1-R_{it}}, \quad (2.6)$$

where $H_{i,t-1}$ is the response history up to time $t-1$ defined as

$$H_{i,t-1} = (y_{i,t-1}, \cdots, y_{i1}, x_i, R_{i,t-1} = 1, \cdots, R_{i1} = 1).$$

Suppose that given the response history, we consider the $P(R_{it} = 1 \mid H_{i,t-1})$ as

$$\begin{aligned} P(R_{it} = 1 \mid H_{i,t-1}) &= \frac{exp(1 + \sum_{j=1}^{q} \alpha_j y_{i,t-j})}{1 + exp(1 + \sum_{j=1}^{q} \alpha_j y_{i,t-j})} \\ &= g_{it}(\alpha \mid y_{i,t-1}, \cdots, y_{i,t-q}), \text{ say,} \end{aligned} \quad (2.7)$$

for $t = 2, \cdots, T$, where $\alpha \equiv (\alpha_1, \cdots, \alpha_q)$ denotes the dependence parameter of the response indicator on the past responses. For simplicity, we will consider $q = 1$ only, in the thesis. Note that as $R_{i1} = 1$ with probability 1, without any loss of generality, we may use $g_{it}(\alpha \mid y_{i,t-1}, \cdots, y_{i,t-q}) = 1$, for t=1.

2.2 Proposed Conditional Probability Models for Incomplete Longitudinal Binary Data

Note that since y_{it} can be available only when $R_{i1} = \cdots = R_{it} = 1$, y_{it} $(t \geq 2)$ conditional on $H_{i,t-1}$, follows a binary model with probability μ^*_{it} (say), where

$$\begin{aligned} \mu^*_{it} &= P(R_{it} = 1, Y_{it} = 1 \mid H_{i,t-1}) \\ &= P(R_{it} = 1 \mid H_{i,t-1}) \Big[P(Y_{it} = 1 \mid R_{it} = 1, H_{i,t-1}) \Big]. \qquad (2.8) \end{aligned}$$

Since it is assumed that the missingness occur in a monotonic pattern, $R_{it} = 1$ may occur only when $R_{i1} = \cdots = R_{i,t-1} = 1$. It therefore follows that

$$P(R_{it} = 1 \mid H_{i,t-1}) = P\left[(\prod_{j=1}^{t} R_{ij} = 1) \mid H_{i,t-1} \right]. \qquad (2.9)$$

Note that when the history $H_{i,t-1}$ is given, the response indicators $R_{i1}, \cdots, R_{it}$ may be considered to be independent. Thus, the probability in (2.9) may be written as

$$\begin{aligned} P(R_{it} = 1 \mid H_{i,t-1}) &= \prod_{j=1}^{t} g_{ij}(\alpha \mid y_{i,j-1}, \cdots, y_{i,j-q}) \\ &= w_{it}(\alpha \mid y_{i,t-1}, \cdots, y_{i1}) \text{ , say,} \qquad (2.10) \end{aligned}$$

where $g_{ij}(\alpha \mid y_{i,j-1}, \cdots, y_{i,j-q})$ is given in (2.7) for $j = 1, \cdots, t$. Further, since the observed repeated binary responses are assumed to follow the AR(1) model (1.26) [or (1.8)], conditional on $H_{i,t-1}$ and $R_{it} = 1$, the binary conditional probability represented by the second term in (2.8) may be written as

$$\begin{aligned} P(Y_{it} = 1 \mid R_{it} = 1, H_{i,t-1}) &= E(Y_{it} \mid R_{it} = 1, H_{i,t-1}) \\ &= \lambda_{i,t|t-1}(\beta, \rho) \\ &= \mu_{it} + \rho(y_{i,t-1} - \mu_{i,t-1}), \end{aligned}$$

by (1.26). Consequently, μ^*_{it} in (2.8) has the formula given by

$$\mu^*_{it} = \lambda_{i,t|t-1}(\beta, \rho) w_{it}(\alpha), \tag{2.11}$$

where $w_{it}(\alpha)$ is used for $w_{it}(\alpha \mid y_{i,t-1}, \cdots, y_{i1})$, for convenience. Note that this conditional probability accommodates both longitudinal correlation structure through $\lambda_{it|t-1}(\beta, \rho)$ and the missingness nature through $w_{it}(\alpha)$.

2.2.1 Weighted response variable

Note that when the longitudinal data is MCAR, it is clear that

$$E[R_{it}(Y_{it} - \mu_{it})] = E(R_{it})E(Y_{it} - \mu_{it}) = 0. \tag{2.12}$$

This unbiasedness property (2.12) for $R_{it}(Y_{it} - \mu_{it})$ however does not hold when the longitudinal data is MAR. This is because when the data is MAR, y_{it} and R_{it} are correlated as both depend on the past history in the longitudinal set-up. Thus under MAR,

$$E[R_{it}(Y_{it} - \mu_{it})] \neq 0. \tag{2.13}$$

But, as by (2.10), $P\left[R_{it} = 1 \mid H_{i,t-1}\right] = w_{it}(\alpha)$, and because conditional on $H_{i,t-1}$, R_{it} and y_{it} are independent, it then follows that

$$\begin{aligned} E\left[\frac{R_{it}}{w_{it}(\alpha)} Y_{it}\right] &= E_{H_{i,t-1}} E\left[\frac{R_{it}}{w_{it}(\alpha)} Y_{it} \mid H_{i,t-1}\right] \\ &= E_{H_{i,t-1}}\left[w^{-1}_{it}(\alpha) E(R_{it} \mid H_{i,t-1}) E(Y_{it} \mid H_{i,t-1})\right] \\ &= E_{H_{i,t-1}}[\lambda_{it|t-1}(\beta, \rho)] \\ &= \mu_{it}. \end{aligned} \tag{2.14}$$

Consequently, in the present incomplete longitudinal data set-up, it is appropriate to use the weighted variable

$$\frac{R_{it}}{w_{it}(\alpha)} y_{it} = \delta_{it}(\alpha) y_{it}$$

instead of the original response variable y_{it} to construct a suitable estimating equation for β involved in μ_{it}. This is equivalent to say by (2.14) that β should be estimated by minimizing the distance function

$$\delta_{it}(\alpha)y_{it} - \mu_{it} \tag{2.15}$$

for all $i = 1, \cdots, K$ and $t = 1, \cdots, T$.

Note that since it is also true that

$$\begin{aligned} E\left[\delta_{it}(\alpha)(Y_{it} - \mu_{it})\right] &= E_{H_{i,t-1}} E\left[\delta_{it}(\alpha)(Y_{it} - \mu_{it}) \mid H_{i,t-1}\right] \\ &= 0, \end{aligned} \tag{2.16}$$

some authors [see Robins, Rotnitzky and Zhao (1995), for example] have minimized the distance function

$$\delta_{it}(\alpha)(y_{it} - \mu_{it}) \tag{2.17}$$

for all $i = 1, \cdots, K$ and $t = 1, \cdots, T$, under the assumption that $H_{i,t-1}$ in $\delta_{it}(\alpha)$ is known. But the unbiasedness truly reflects only when the expectation over history is taken.

Note however that the distance functions in (2.15) and (2.17) are not the same, i.e.,

$$\delta_{it}(\alpha)y_{it} - \mu_{it} \neq \delta_{it}(\alpha)y_{it} - \delta_{it}(\alpha)\mu_{it}, \tag{2.18}$$

this is because $\delta_{it}(\alpha)\mu_{it} \neq \mu_{it}$ even though their conditional expectations are the same, i.e.,

$$\begin{aligned} E\left[\delta_{it}(\alpha)y_{it} - \mu_{it} \mid H_{i,t-1}\right] &= E\left[\delta_{it}(\alpha)y_{it} - \delta_{it}(\alpha)\mu_{it} \mid H_{i,t-1}\right] \\ &= \lambda_{it|t-1}(\beta, \rho) - \mu_{it}, \end{aligned} \tag{2.19}$$

implying that their unconditional expectations are also the same as

$$E\left[\delta_{it}(\alpha)y_{it} - \mu_{it}\right] = E\left[\delta_{it}(\alpha)y_{it} - \delta_{it}(\alpha)\mu_{it}\right] = 0. \tag{2.20}$$

Further, (2.19) reveals that it may be much better to adopt conditional inference as opposed to unconditional inferences for β. Thus conditional on $H_{i,t-1}$, one may minimize the distance function

$$\delta_{it}(\alpha)y_{it} - \lambda_{it|t-1}(\beta, \rho) \tag{2.21}$$

for $i = 1, \cdots, K$ and $t = 1, \cdots, T$. Note that the conditional (on $H_{i,t-1}$) expectation of this distance function (2.21) is zero. That is,

$$E\left[\delta_{it}(\alpha)y_{it} - \lambda_{it|t-1}(\beta, \rho) \mid H_{i,t-1}\right] = 0. \tag{2.22}$$

This implies that as opposed to considering a distance function which needs to be unbiased for zero unconditionally, it is enough to consider the conditional (on $H_{i,t-1}$) distance function (2.21) for the unbiased estimation of β. If one however would like to use a distance function with its unconditional expectation as zero, it would be better to use the distance function $\delta_{it}(\alpha)y_{it} - \mu_{it}$ (2.15) instead of $\delta_{it}(\alpha)y_{it} - \delta_{it}(\alpha)\mu_{it}$ (2.17), the later function being exploited in the existing literature. Note that it is also recognized in the literature that minimization of the distance function $\delta_{it}(\alpha)\left(y_{it} - \mu_{it}\right)$ (2.17) may still produce biased estimate for the regression effects β. Some authors such as Rotnitzky, Robins and Scharfstein (1998) attempted to exploit an inverse probability of censoring weighted (IPCW) type distance function given by

$$\delta_{it}(\alpha)\left(y_{it} - \mu_{it}\right) + a^*_{it}$$

for further bias correction, a^*_{it} being an augmented function [see eqn. (11), p. 1327 in Rotnitzky et al. (1998)]. This correction appears to be complicated. Furthermore, it is demonstrated in the thesis that utilization of the simpler but correct distance function (2.15) may be enough to obtain unbiased estimates. Thus, we do not discuss any further about the IPCW augmentation approach in the present thesis.

Note that Yi and Cook (2002) have also dealt with the inferences for the regression effects in incomplete longitudinal data set-up. Similarly to Robins et al. (1995), these authors have exploited the basic distance function $\delta_{it}(\alpha)(y_{it} - \mu_{it})$ (2.17) in constructing their estimating equations for β [see Yi and Cook (2002), eqn. (5), p.1074], whereas we suggest to exploit the distance function $\delta_{it}(\alpha)y_{it}-\mu_{it}$ (2.15) for the reasons explained above. Furthermore, Yi and Cook (2002) modelled the association among the longitudinal data through odds ratios which are mainly suitable for binary data [Yi and Cook (2002), eqn. (1)] as opposed to the count data. Furthermore, since the odds ratios are unknown, they are estimated by using an additional (on top of the regression relation between the main variables y_{it} and covariates x_{it}) regression relationship between odds ratios and certain new covariates. This relationship appears to be arbitrary. See Sutradhar and Kovacevic (2000) [see also Williamson, Kim and Lipsitz (1995)] for more discussion on this.

In what follows, we shall let

$$\tilde{y}_{it} = \delta_{it}(\alpha)y_{it} = R_{it}y_{it}/w_{it}(\alpha) \tag{2.23}$$

be the weighted response variable which is involved in the unconditional distance function (2.15), as well as in the conditional distance function (2.21).

2.2.2 Conditional first and second order moments of the weighted response variable (2.23) under incomplete MAR model

a. Conditional mean

It follows from (2.14) or (2.19) that

$$\begin{aligned} E(\tilde{Y}_{it} \mid H_{i,t-1}) &= \lambda_{it|t-1}(\beta, \rho) \\ &= \mu_{it} + \rho(y_{i,t-1} - \mu_{i,t-1}). \end{aligned} \tag{2.24}$$

b. Conditional variance and covariance

Conditional on the history $H_{i,t-1}$, one may write

$$E\left[\tilde{Y}^2_{it} \mid H_{i,t-1}\right] = E\left[\frac{R^2_{it}Y^2_{it}}{w^2_{it}(\alpha)} \mid H_{i,t-1}\right]$$

$$\begin{aligned} &= E\left[\frac{R_{it}Y_{it}}{w_{it}^2(\alpha)} \mid H_{i,t-1}\right] \\ &= E\left[\frac{\tilde{Y}_{it}}{w_{it}(\alpha)} \mid H_{i,t-1}\right] \\ &= \frac{\lambda_{it|t-1}(\beta,\rho)}{w_{it}(\alpha)}, \end{aligned}$$

by (2.24). This yields the conditional variance of $\tilde{y}_{it}$ as

$$\begin{aligned} var(\tilde{Y}_{it} \mid H_{i,t-1}) &= E(\tilde{Y}_{it}^2 \mid H_{i,t-1}) - \left[E(\tilde{Y}_{it} \mid H_{i,t-1})\right]^2 \\ &= \frac{\lambda_{it|t-1}(\beta,\rho)}{w_{it}(\alpha)} - \left[\lambda_{it|t-1}(\beta,\rho)\right]^2 \\ &= \lambda_{it|t-1}(\beta,\rho)\left[w_{it}^{-1}(\alpha) - \lambda_{it|t-1}(\beta,\rho)\right]. \qquad (2.25) \end{aligned}$$

By similar calculations, we may show that for $l = max(t,t^{'})$, $t \neq t^{'}$, $t,t^{'} = 1,\cdots,T$,

$$cov\left[\tilde{Y}_{it}, \tilde{Y}_{it^{'}} \mid H_{i,l-1}\right] = 0. \qquad (2.26)$$

This is because, for $t > t^{'}$

$$\begin{aligned} cov\left[\tilde{Y}_{it}, \tilde{Y}_{it^{'}} \mid H_{i,t-1}\right] &= E\left[\tilde{Y}_{it}\tilde{Y}_{it^{'}} \mid H_{i,t-1}\right] - E(\tilde{Y}_{it} \mid H_{i,t-1})E(\tilde{Y}_{it^{'}} \mid H_{i,t-1}) \\ &= \tilde{Y}_{it^{'}}E\left(\tilde{Y}_{it} \mid H_{i,t-1}\right) - \tilde{Y}_{it^{'}}E\left(\tilde{Y}_{it} \mid H_{i,t-1}\right) \\ &= 0. \end{aligned}$$

Note that the conditional variances (2.25) and the covariances (2.26) will be exploited in Section 2.3 to construct a conditional weighted generalized quasilikelihood (CWGQL) estimating equation for β.

2.2.3 Unconditional first and second order moments of the weighted response variable (2.23) under incomplete MAR model

a. Unconditional mean

It follows from (2.14) that

$$\begin{aligned} E(\tilde{Y}_{it}) &= E\left[\frac{R_{it}Y_{it}}{w_{it}(\alpha)}\right] \\ &= E_{H_{i,t-1}}E\left[\frac{R_{it}}{w_{it}(\alpha)}Y_{it} \mid H_{i,t-1}\right] \\ &= \mu_{it}. \end{aligned}$$

b. Unconditional variance and covariance

Note that in general it is not easy to compute the unconditional variances and the covariances of the weighted response variables (2.23). The complexity in computing these variances and covariances may be demonstrated by considering the computations of $V(\tilde{Y}_{i3})$ and $cov(\tilde{Y}_{i2}, \tilde{Y}_{i4})$, for example.

(i) Computation of $V(\tilde{Y}_{i3})$

Write

$$\begin{aligned} V(\tilde{Y}_{i3}) &= E(\tilde{Y}_{i3}^2) - \left[E(\tilde{Y}_{i3})\right]^2 \\ &= E(\tilde{Y}_{i3}^2) - \mu_{i3}^2, \end{aligned}$$

by (2.14). However, as we demonstrate below, computing $E(\tilde{Y}_{i3}^2)$ can be cumbersome. The steps are as follows for this computation. First, we write

$$E\left[\tilde{Y}_{i3}^2\right] = E\left[\frac{\tilde{Y}_{i3}}{w_{i3}(\alpha \mid Y_{i2}, Y_{i1})}\right]$$

$$
\begin{aligned}
&= E_{H_{i2}} E\left[\frac{\tilde{Y}_{i3}}{w_{i3}(\alpha \mid Y_{i2}, Y_{i1})} \mid H_{i2}\right] \\
&= E_{Y_{i1},Y_{i2}}\left[w_{i3}^{-1}(\alpha \mid Y_{i2}, Y_{i1})\lambda_{i3|2}(Y_{i2})\right],
\end{aligned}
$$

by (2.24). Note that, since $\lambda_{it|t-1}(\beta, \rho) = \mu_{it} + \rho(y_{i,t-1} - \mu_{i,t-1})$ is a function of $y_{i,t-1}$, we use $\lambda_{it|t-1}(y_{i,t-1})$, for convenience, to represent $\lambda_{it|t-1}(\beta, \rho)$. Now we take the expectation over y_{i2} given y_{i1}. Thus,

$$
\begin{aligned}
E\left[\tilde{Y}_{i3}^{2}\right] &= E_{Y_{i1}}\left[E_{Y_{i2}}\left\{w_{i3}^{-1}(\alpha \mid Y_{i2}, Y_{i1})\lambda_{i3}(Y_{i2}) \mid Y_{i1}\right\}\right] \\
&= E_{Y_{i1}}\left[w_{i3}^{-1}(\alpha \mid 1, Y_{i1})\lambda_{i3}(1)P(Y_{i2} = 1 \mid Y_{i1})\right. \\
&\quad \left. + w_{i3}^{-1}(\alpha \mid 0, Y_{i1})\lambda_{i3|2}(0)P(Y_{i2} = 0 \mid Y_{i1})\right]. \qquad (2.27)
\end{aligned}
$$

Next, we take the expectation over y_{i1}. This gives

$$
\begin{aligned}
E\left[\tilde{Y}_{i3}^{2}\right] &= \frac{\lambda_{i3}(1)\lambda_{i2}(1)\mu_{i1}}{w_{i3}(\alpha \mid 1, 1)} + \frac{\lambda_{i3}(0)[1 - \lambda_{i2}(1)]\mu_{i1}}{w_{i3}(\alpha \mid 0, 1)} \\
&\quad + \frac{\lambda_{i3}(1)\lambda_{i2}(0)(1 - \mu_{i1})}{w_{i3}(\alpha \mid 1, 0)} + \frac{\lambda_{i3}(0)[1 - \lambda_{i2}(0)][1 - \mu_{i1}]}{w_{i3}(\alpha \mid 0, 0)}, \qquad (2.28)
\end{aligned}
$$

with $\lambda_{i2}(0) \equiv \lambda_{i2}(y_{i,1} = 0) = \mu_{i2} - \rho\mu_{i1}$, for example, and by (2.7) and (2.10)

$$
\begin{aligned}
w_{i3}(\alpha \mid 1, 0) &= g_{i3}(\alpha \mid y_{i2} = 1)g_{i2}(\alpha \mid y_{i1} = 0) \\
&= \frac{exp(1 + \alpha)}{1 + exp(1 + \alpha)}\left[\frac{exp(1)}{1 + exp(1)}\right]. \qquad (2.29)
\end{aligned}
$$

Thus, it is clear that $E\left(\tilde{Y}_{i3}^{2}\right)$ given by (2.28) is cumbersome to compute.

(ii) Computation of $cov[\tilde{\mathbf{Y}}_{i2}, \tilde{\mathbf{Y}}_{i4}]$

To compute this covariance, we write

$$
\begin{aligned}
cov[\tilde{Y}_{i2}, \tilde{Y}_{i4}] &= E(\tilde{Y}_{i2}\tilde{Y}_{i4}) - E(\tilde{Y}_{i2})E(\tilde{Y}_{i4}) \\
&= E\left[\frac{R_{i2}Y_{i2}}{w_{i2}(\alpha \mid Y_{i1})}\left\{\frac{R_{i4}Y_{i4}}{w_{i4}(\alpha \mid Y_{i3}, Y_{i2}, Y_{i1})}\right\}\right] - \mu_{i2}\mu_{i4} \qquad (2.30)
\end{aligned}
$$

Note that the computation for the first part in (2.30) can be done as follows

$$E\left[\frac{R_{i2}Y_{i2}}{w_{i2}(\alpha \mid Y_{i1})}\left\{\frac{R_{i4}Y_{i4}}{w_{i4}(\alpha \mid Y_{i3},Y_{i2},Y_{i1})}\right\}\right] = E_{Y_{i3},Y_{i2},Y_{i1}}E_{R_{i4},Y_{i4}}\left[\frac{Y_{i2}}{w_{i2}(\alpha \mid Y_{i1})}\times\right.$$

$$\left.\frac{R_{i4}Y_{i4}}{w_{i4}(\alpha \mid Y_{i3},Y_{i2},Y_{i1})} \mid R_{i3}=1, R_{i2}=1, Y_{i3}, Y_{i2}, Y_{i1}\right]. \quad (2.31)$$

Since $E_{R_{i4}Y_{i4}}[R_{i4}Y_{i4}/w_{i4}(\alpha \mid Y_{i3},Y_{i2},Y_{i1}) \mid H_{i3}] = \lambda_{i4}(Y_{i3})$ by (2.24), by similar arguments as in (2.27) and (2.28), we compute the expectation in (2.31) as

$$E\left[\frac{R_{i2}Y_{i2}}{w_{i2}(\alpha \mid Y_{i1})}\left\{\frac{R_{i4}Y_{i4}}{w_{i4}(\alpha \mid Y_{i3},Y_{i2},Y_{i1})}\right\}\right] = E_{Y_{i3},Y_{i2},Y_{i1}}\left[\frac{Y_{i2}}{w_{i2}(\alpha \mid Y_{i1})}\lambda_{i4}(Y_{i3})\right]$$

$$= [\rho\lambda_{i3}(1)+\lambda_{i4}(0)]\left[\frac{\mu_{i1}\lambda_{i2}(1)}{w_{i2}(\alpha \mid 1)}+\frac{(1-\mu_{i1})\lambda_{i2}(0)}{w_{i2}(\alpha \mid 0)}\right], \quad (2.32)$$

which is also relatively cumbersome to compute. In fact the computation of $E(\tilde{Y}_{it}^2)$ for t larger than 3 would be much more complex. By the same token, the computations for the higher lagged covariances will also be complicated. This computational difficulty seems to be a major drawback for the exploitation of the unconditional distance function (2.15) for the estimation of β.

Note that even though the unconditional distance function, more specifically, the distance function in (2.17) has been used in the existing literature [see Robins et al. (1995)] in order to write a so-called generalized estimating equation (GEE), the variance-covariance matrix of the unconditional distance functions was neither computed nor used. Instead, the authors have used an arbitrary 'working' covariance matrix which was not clearly defined either in terms of the longitudinal correlations or missingness or both. In the thesis, we will not follow this unconditional approach any further, except that our conditional inferences will be compared numerically with the corresponding unconditional inferences where estimating equations are constructed using arbitrary covariance matrix.

2.3 Proposed Conditional Weighted Generalized Quasilikelihood (CWGQL) Estimating Equations

Recall from (2.24) that the expectation of the weighted variable $\tilde{y}_{it} = R_{it}y_{it}/w_{it}(\alpha)$ conditional on $H_{i,t-1}$ is given by

$$E(\tilde{Y}_{it} \mid H_{i,t-1}) = \lambda_{it|t-1}(\beta, \rho).$$

Also the formulas for the conditional variance, $V(\tilde{Y}_{it} \mid H_{i,t-1})$ and the conditional auto-covariance of lag $|t - t'|$ are given by (2.25) and (2.26), respectively.

Consider

$$\tilde{y}_i = [\tilde{y}_{i1}, \cdots, \tilde{y}_{iT_i}]'$$

be the $T_i \times 1$ vector of available weighted responses. Note that, under the incomplete data set-up, we assume that $\sum_{t=1}^{T} R_{it} = T_i$ $(1 \leq T_i \leq T)$ repeated responses are available implying $T - T_i$ responses are missing at random for the i-th subject. Let

$$\lambda_i = [\lambda_{i1}, \lambda_{i2|1}(\beta, \rho), \cdots, \lambda_{iT_i|T_i-1}(\beta, \rho)]'$$

be the conditional mean vector of $\tilde{y}_i$ with

$$\lambda_{i1} = E(\tilde{Y}_{i1}) = E(Y_{i1}) = \mu_{i1} \text{ (as } R_{i1} = 1 \text{ and } w_{i1}(\alpha) = 1)$$

and

$$\lambda_{it|t-1}(\beta, \rho) = E(\tilde{Y}_{it} \mid H_{i,t-1}) = \mu_{it} - \rho(y_{i,t-1} - \mu_{i,t-1}),$$

where $\mu_{it} = exp(x_{it}'\beta)/[1 + exp(x_{it}'\beta)]$, for all $t = 1, \cdots, T_i$.

Next, suppose that $\tilde{\Sigma}_{iw}$ is the true conditional covariance matrix of $\tilde{y}_i$. That is

$$\tilde{\Sigma}_{iw} = cov(\tilde{Y}_i) = (\tilde{\sigma}_{itt'}), \tag{2.33}$$

where the formulas for $\tilde{\sigma}_{itt'}$ for $t = t'$ by (2.25) is written as

$$\begin{aligned} \tilde{\sigma}_{itt} &= V(\tilde{Y}_{it} \mid H_{i,t-1}) \\ &= w_{it}^{-1}(\alpha)\lambda_{i,t|t-1}(\beta, \rho) - \lambda_{i,t|t-1}^2(\beta, \rho), \end{aligned}$$

and for $t \neq t'$, by (2.26), is written as

$$\tilde{\sigma}_{itt'} = 0.$$

Note that for $t = 1$, $V(Y_{i1}) = \mu_{i1}(1 - \mu_{i1})$, which is the same as $\tilde{\sigma}_{i11}$. This is because for $t = 1$, $w_{i1}(\alpha) = 1$ and $\lambda_{i1} = \mu_{i1}$.

Now, in the fashion similar to that of Sutradhar (2003, section 3), we write the GQL estimating equation in terms of the conditional distance functions, as

$$\sum_{i=1}^{K} \frac{\partial \lambda_i'}{\partial \beta} [\tilde{\Sigma}_{iw}]^{-1} [\tilde{y}_i - \lambda_i] = 0. \tag{2.34}$$

Note that this equation in (2.34) may be referred to as the conditional weighted GQL (CWGQL) estimating equation. This is because, (2.34) exploits the conditional mean vector λ_i as well as conditional weighted covariance matrix $\tilde{\Sigma}_{iw}$, whereas the ordinary QL approach [see Wedderburn (1974) and McCullagh (1983), for example] exploits the means and variances of the independent data.

Next, the formulas for the elements of the $p \times T_i$ derivative matrix, $\partial \lambda_i' / \partial \beta$ in (2.34), may be obtained by computing the derivative of a general element of λ_i with respect to the s-th $(s = 1, \cdots, p)$ element of β. These formulas are

$$\frac{\partial \lambda_{it|t-1}(\beta, \rho)}{\partial \beta_s} = \begin{cases} x_{i1s}\mu_{i1}(1 - \mu_{i1}), & \text{for } t = 1 \\ x_{its}\mu_{it}(1 - \mu_{it}) - \rho x_{i,t-1,s}\mu_{i,t-1}(1 - \mu_{i,t-1}), & \text{for } t > 1 \end{cases}, \tag{2.35}$$

Note that the CWGQL estimating equation (2.34) is constructed by exploiting the distance function $\tilde{y}_{it} - \lambda_{it|t-1}(\beta, \rho)$ which is unbiased for zero, conditional on the history $H_{i,t-1}$ as given by (2.22). Further note that, the CWGQL estimating equation uses the conditional covariance matrix $\tilde{\Sigma}_{iw}$ which accommodates both the longitudinal correlation structure and the missing mechanism. The use of the proper weight matrix $\tilde{\Sigma}_{iw}$ for $\tilde{y}_i$ makes the CWGQL estimating equation (2.34) as an efficient estimating equation for the regression effects β. This equation is also easy to compute.

Further note that we have considered the monotonic response pattern in the present thesis. But, if the responses are assumed to be missing intermittently, it may not be easy to implement the conditional approach, which is, however, beyond the scope of the present thesis.

2.3.1 Estimation of ρ

It is clear from the last sub-section that solution for β by CWGQL (2.34) approach requires the longitudinal correlation parameter ρ to be known. But, as ρ is unknown in practice, we provide a formula for a consistent estimator of this parameter by using the well-known method of moments. For the purpose, we first consider a Pearsonian type of correlation given by

$$V = \frac{\sum_{i=1}^{K}\sum_{t=2}^{T} R_{it}R_{i,t-1}y^*_{it}y^*_{i,t-1}}{\sum_{i=1}^{K}\sum_{t=1}^{T} R_{it}y^{*2}_{it} / \sum_{i=1}^{K}\sum_{t=1}^{T} R_{it}} = \frac{a}{b},$$

where $y^*_{it} = (y_{it} - \mu_{it})/\sqrt{(\sigma_{i,tt})}$ with $\sigma_{i,tt} = \mu_{it}(1-\mu_{it})$. Note that the standardized variable y^*_{it} is constructed based on complete or unweighted random variable y_{it}. Since $E(a) = \rho \sum_{i=1}^{K}\sum_{t=2}^{T} R_{it}R_{i,t-1}\sqrt{\sigma_{i,t-1,t-1}/\sigma_{i,tt}}$ and $E(b) = 1$, we obtain an approximate moment estimator of ρ given by

$$\hat{\rho}_{MM} = \frac{\sum_{i=1}^{K}\sum_{t=2}^{T} R_{it}R_{i,t-1}y^*_{it}y^*_{i,t-1}}{\sum_{i=1}^{K}\sum_{t=1}^{T} R_{it}y^{*2}_{it}} \frac{\sum_{i=1}^{K}\sum_{t=1}^{T} R_{it}}{\sum_{i=1}^{K}\sum_{t=2}^{T} \left[R_{it}R_{i,t-1}\sigma_{i,t-1,t-1}/\sigma_{i,tt}\right]^{1/2}}. \qquad (2.36)$$

Note that in (2.36), $\hat{\rho}_{MM}$ denote the method of moments (MM) estimator of ρ suitable for complete data, implying that this MM estimator is obtained by ignoring the missing mechanism.

We also consider a modification of the above formula (2.36), where we use the weighted variable $\tilde{y}_{it} = w^{-1}_{it}(\alpha)y_{it}$ in place of y_{it}, where $\tilde{y}_{it}$ is formulated by accommodating the weights due to the non-response mechanism. To be specific, we define $z^*_{it} = (\tilde{y}_{it} - \mu_{it})/\sqrt{(\sigma_{i,tt})}$. This gives an estimator of ρ similar to (2.36), which we refer to as the weighted method of moment (WMM) estimator and denote it by $\hat{\rho}_{WMM}$. To be specific, the formula for $\hat{\rho}_{WMM}$ is given by

$$\hat{\rho}_{WMM} = \frac{\sum_{i=1}^{K}\sum_{t=2}^{T} R_{it}R_{i,t-1}z^*_{it}z^*_{i,t-1}}{\sum_{i=1}^{K}\sum_{t=1}^{T} R_{it}z^{*2}_{it}} \frac{\sum_{i=1}^{K}\sum_{t=1}^{T} R_{it}}{\sum_{i=1}^{K}\sum_{t=2}^{T} \left[R_{it}R_{i,t-1}\sigma_{i,t-1,t-1}/\sigma_{i,tt}\right]^{1/2}}. \qquad (2.37)$$

Note that in (2.37) inverse weights are used in a multiplicative (i.e. independence assumption based) form for a pair of standardized responses. One could however attempt to model the correlation of joint indication of two responses conditional on the common past, which would have allowed to write a joint weight for paired responses in (2.37). We however have chosen the multiplicative form for simplicity.

2.4 Existing Unconditional Weighted Generalized Estimating Equations (UWGEE)

Even though $\delta_{it}(\alpha)y_{it} - \mu_{it}$ in (2.15) is a proper unconditional distance function, Robins et al. (1995) however used a different unconditional distance function namely $\delta_{it}(\alpha)(y_{it} - \mu_{it})$ (2.17) to construct an unconditional weighted GEE (UWGEE) given by

$$\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta}[\Sigma^*_{i(comp)}(\alpha^*)]^{-1}\Delta_i(\alpha)[y_i - \mu_i] = 0, \tag{2.38}$$

[see also Paik (1977, eq.(1), p. 1321)], where $\Delta_i(\alpha) = diag[\delta_{i1}(\alpha), \cdots, \delta_{iT}(\alpha)]$ with $\delta_{it}(\alpha) = R_{it}/w_{it}(\alpha)$ (such that $R_{i1} = \cdots = R_{iT_i} = 1, R_{i,T_i+1} = \cdots = R_{iT} = 0$ for $T_i \leq T$) and $\Sigma^*_{i(comp)}(\alpha^*) = A_i^{1/2}C_i^*(\alpha^*)A_i^{1/2}$ is a 'working' covariance matrix constructed for the complete data case even though the actual covariance matrix under the incomplete longitudinal set-up would be different. More specifically, this 'working' covariance matrix is constructed by ignoring both (i) missing mechanism and the (ii) true correlation structure of the data. Thus, these authors have used $A_i = diag[V(Y_{i1}), \cdots, V(Y_{iT})]$ with $V(Y_{it}) = \mu_{it}(1 - \mu_{it})$ as the variance of y_{it} formulated by pretending that the data were complete. Similarly, the correlation matrix $C_i^*(\alpha^*)$ has also been constructed by pretending that the data were complete. Furthermore, similar to Liang and Zeger (1986), they suggested to use a 'working' correlation structure based on 'working' correlation α^*. But, as Sutradhar and Das (1999) [see also Sutradhar (2003)] pointed out, this 'working' correlation approach may produce less efficient estimates than the $\alpha^* = 0$ (independence) case. For incomplete longitudinal data, using such $C_i^*(\alpha^*)$ would, naturally, have more effect on the estimates. Nevertheless, we incorporate this UWGEE approach in our simulation study in Section 2.6 to examine its comparative behavior with that of the proposed CWGQL approach discussed in the last section.

In (2.38), the elements of the derivative matrix $\partial\mu_i'/\partial\beta$ may be obtained by computing the derivative of the t-th element of μ_i with respect to the s-th $(s = 1, \cdots, p)$ element of β. To be specific

$$\frac{\partial\mu_{it}}{\partial\beta_s} = x_{its}\mu_{it}(1 - \mu_{it}). \tag{2.39}$$

2.5 A Modified UWGEE (MUWGEE) Approach

Since the modified unconditional distance function $\{\delta_{it}(\alpha)y_{it} - \mu_{it}\}$ in (2.15) is more appealing than the existing unconditional distance function $\delta_{it}(\alpha)(y_{it} - \mu_{it})$, we now use the former distance function and construct a UWGEE similar to (2.38). This modified UWGEE (MUWGEE) is given by

$$\sum_{i=1}^{K} \frac{\partial\mu_i'}{\partial\beta}[\Sigma_{i(comp)}^*(\rho)]^{-1}[\Delta_i(\alpha)y_i - \mu_i] = 0. \tag{2.40}$$

We have included this MUWGEE approach as well in our simulation study.

2.6 Empirical Study

2.6.1 Generation of the data

To generate longitudinal binary data subject to MAR, we follow the conditional probability model introduced in Section 2.2. To be specific, we follow equation (2.11), to generate such incomplete binary data. As far as the simulation design is concerned, we use:

$$K = 100,\ T = 4,\ p = 2,\ q = 1,\ \alpha = 4 \text{ and/or } 1,$$

$$\rho = 0,\ 0.2,\ 0.4,\ 0.6 \text{ and } 0.8, \text{ and } \beta_1 = \beta_2 = 0;$$

and the time dependent covariates

$$x_{it1} = \begin{cases} \frac{1}{2} & \text{for } i = 1, \cdots, \frac{K}{4};\ t = 1, 2 \\ 0 & \text{for } i = 1, \cdots, \frac{K}{4};\ t = 3, 4 \\ -\frac{1}{2} & \text{for } i = \frac{K}{4} + 1, \cdots, \frac{3K}{4};\ t = 1 \\ 0 & \text{for } i = \frac{K}{4} + 1, \cdots, \frac{3K}{4};\ t = 2, 3 \\ \frac{1}{2} & \text{for } i = \frac{K}{4} + 1, \cdots, \frac{3K}{4};\ t = 4 \\ \frac{t}{2T} & \text{for } i = \frac{3K}{4} + 1, \cdots, K;\ t = 1, \cdots, 4 \end{cases}$$

and

$$x_{it2} = \begin{cases} \frac{t-2.5}{2T} & \text{for } i = 1, \cdots, \frac{K}{2};\ t = 1, \cdots, 4 \\ 0 & \text{for } i = \frac{K}{2} + 1, \cdots, K;\ t = 1, 2 \\ \frac{1}{2} & \text{for } i = \frac{K}{2} + 1, \cdots, K;\ t = 3, 4 \end{cases}$$

respectively. Note that these covariates are chosen to reflect their certain categorical nature over time. For convenience, we summarize the data generation in the following steps:

Step 1: Under the assumption that all individuals provide their first response, i.e., $R_{i1} = 1$ for all $i = 1, \cdots, K$, we generate y_{i1} as $y_{i1} \sim bin(\mu_{i1})$ with $\mu_{i1} = exp(x_{i1}'\beta)/[1 + exp(x_{i1}'\beta)]$, where $x_{i1} = (x_{i11}, x_{i12})'$ and $\beta = (\beta_1, \beta_2)'$.

Step 2: Generate the second response indicator R_{i2} as $R_{i2} \sim bin[g_{i2}(\alpha \mid y_{i1})]$, with $g_{i2}(\alpha \mid y_{i1}) = exp(1 + \alpha y_{i1})/[1 + exp(1 + \alpha y_{i1})]$. Note that the response indicator R_{i2} is generated as a function of the previous response y_{i1}, satisfying the definition of MAR mechanism.

Step 3: If $R_{i2} = 0$, stop the process. This implies that no more responses beyond y_{i1} will be available, for the i-th ($i = 1, \cdots, K$) individual. However, if $R_{i2} = 1$, generate y_{i2} as $y_{i2} \sim bin[\lambda_{i2|1}(\beta, \rho) w_{i2}(\alpha)]$ following (2.11) and return to step 2 to generate R_{i3}. Recall that $\lambda_{i2|1}(\beta, \rho) = \mu_{i2} + \rho(y_{i1} - \mu_{i1})$ is the AR(1) based regression function, ρ being the longitudinal correlation between y_{i1} and y_{i2}. Note that in the existing literature this stochastic correlation structure was not used so far in any studies for incomplete longitudinal data analysis. For example, Robins et al. (1995, §5, p. 113) generated y_{it} following a linear relationship in t, which does not accommodate the longitudinal correlation between y_{it} $(t = 2, \cdots, T)$ and the past responses, namely $y_{i,t-1}, \cdots, y_{i1}$. Further note that the proposed data generation mechanism also accommodates the missing at random model through $w_{i2}(\alpha) = w_{i2}(\alpha \mid y_{i1})$.

Note that if y_{i2} were generated following the conditional probability

$$\lambda_{i2|1}(\beta, \rho) = P(Y_{i2} = 1 \mid R_{i2} = 1, y_{i1}),$$

one could then have avoided the effect of the missing mechanism on the observed responses. But, this avoidance does not appear to be sound in the present incomplete longitudinal set up.

Step 4: Follow steps 2 and 3 and generate $y_{i1}, \cdots, y_{iT_i}$ for $T_i \leq T$. Note that these T_i responses become available only when $R_{i1} = \cdots = R_{iT_i} = 1$.

2.6.2 Performance of the estimators for known ρ

Our main purpose is to examine the performances of various existing and proposed approaches in estimating the regression effects when longitudinal data subject to MAR are generated as in the last section. Note that according to the MAR mechanism, the response availability of an individual at a given time point depends on the previous responses of the individual. As mentioned in Step 2 under data generation, α denotes the effect of the previous responses on such response availability. Furthermore, if there is an affirmative response indication, the actual response is then generated by maintaining a longitudinal correlation structure among the available responses. As indicated in Step 3, ρ is an index for such longitudinal correlations. In the present simulation study, we assume that α is known such as $\alpha = 4$ or 1. Note that as we demonstrate in Section 2.6.4, $\alpha = 4$ represents approximately 9% missing responses, whereas $\alpha = 1$ represents approximately 14% missing responses. As far as the longitudinal correlations are concerned, here we consider that ρ is known and examine the performances of the existing as well as proposed approaches in estimating β. The case for unknown ρ is discussed in Section 2.6.3.

(a) Performance of the UWGEE approach

Based on 1000 simulations, the simulated means (SM), simulated standard errors (SSE) and simulated mean squared errors (SMSE) of the estimates of β_1 and β_2 based on the UWGEE approach are given in Table 2.1. In the same table, we also report the estimates of regression parameters by using a naive unconditional weighted method of moments (UWMM) approach, where this UWMM equation is given by

$$\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} \Delta_i(\alpha)(y_i - \mu_i) = 0. \tag{2.41}$$

Note that UWMM equation in (2.41) is obtained simply by avoiding the 'working' covariance matrix $\Sigma^*_{i(comp)}(\alpha^*)$ in (2.38). This lead (2.38) to be a suitable method of moments equation. Further note that this UWMM was also suggested by Robins et al. (1995, eqn. 10, p.109). As far as the UWGEE approach is concerned, we also provide the regression estimates when $C_i^*(\alpha)$ in (2.38) is computed based on certain mis-specified such as independent (I), MA(1) and equicorrelation (EQC) structures, whereas the responses are generated in Section 2.6.1 following the AR(1) type binary model.

For known $\alpha = 4$ and selected known values of $\rho = 0.0$, 0.2, 0.4, 0.6 and 0.8, we now compute the UWGEE (2.38) estimates of β_1 and β_2. Note that in general, when longitudinal correlation is taken into account, the results in Table 2.1 show that the UWGEE as well as the naive UWMM approaches of Robins et al. (1995) performs poorly. For example, for $\rho = 0.6$, the estimates of β_1 and β_2 under the UWMM approach are found to be -0.250 and -0.868, for true $\beta_1 = 0$ and $\beta_2 = 0$, respectively. The simulated standard errors (SSE) also appear to be large in general, yielding large

Table 2.1: Simulated mean (SM), simulated standard error (SSE) and simulated mean squared error (SMSE) for the UWMM and UWGEE based estimates [Robins et al.(1995)] under binary data with $\beta_1 = \beta_2 = 0.0$, $\alpha = 4$ and selected known values of ρ, based on 1000 simulations

		UWMM		UWGEE							
				True		Working correlation structures					
				AR(1)		(I)		(Eq)		(MA(1))	
ρ	Statistic	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
0.0	SM	-0.365	-1.269	-0.365	-1.254	-0.365	-1.254	-0.365	-1.254	-0.365	-1.254
	SSE	0.378	0.567	0.376	0.557	0.376	0.557	0.376	0.557	0.376	0.557
	SMSE	0.276	1.933	0.275	1.883	0.275	1.883	0.275	1.883	0.275	1.883
0.2	SM	-0.295	-1.250	-0.339	-1.308	-0.296	-1.235	-0.260	-1.526	-0.352	-1.275
	SSE	0.378	0.577	0.365	0.553	0.377	0.570	0.361	0.552	0.367	0.556
	SMSE	0.230	1.895	0.248	2.017	0.230	1.851	0.198	2.634	0.259	1.935
0.4	SM	-0.266	-1.137	-0.377	-1.278	-0.267	-1.125	-0.239	-1.661	-0.440	-1.128
	SSE	0.390	0.613	0.353	0.544	0.388	0.605	0.358	0.564	0.365	0.561
	SMSE	0.223	1.669	0.267	1.930	0.222	1.632	0.186	3.079	0.327	1.588
0.6	SM	-0.250	-0.868	-0.413	-1.146	-0.250	-0.864	-0.212	-1.655	–	–
	SSE	0.387	0.604	0.317	0.475	0.386	0.605	0.344	0.516	–	–
	SMSE	0.212	1.119	0.271	1.538	0.211	1.113	0.163	3.007	–	–
0.8	SM	-0.134	-0.544	-0.403	-0.948	-0.134	-0.543	-0.154	-1.523	–	–
	SSE	0.378	0.619	0.259	0.331	0.378	0.621	0.283	0.391	–	–
	SMSE	0.161	0.679	0.230	1.008	0.161	0.875	0.104	2.472	–	–

simulated mean squared errors (SMSE) 0.212 and 1.119, respectively. The results in Table 2.1 also exhibit that the UWGEE (2.38) approach does not appear to improve the estimates over the UWMM (2.41) approach. In fact in some cases, UWGEE (2.38) approach performs worse, yielding much more biased estimates, as compared to the UWMM (2.41) approach. For example, results based on 'working' MA(1) correlation structure show that for $\rho = 0.4$, $\hat{\beta} = (-0.440, -1.128)'$ under UWGEE approach which is much more biased than that of corresponding UWMM estimates $\hat{\beta} = (-0.266, -1.137)'$. This is not surprising as the UWGEE (2.38) is constructed by inserting a weight matrix in UWMM (2.41), which is the inverse of a 'working' covariance matrix for y_i, instead of a 'working' covariance matrix for the weighted variable $\Delta_i(\alpha)(y_i - \mu_i)$. Note that it does not however imply that $\Delta_i(\alpha)(y_i - \mu_i)$ is an appropriate distance function one should minimize to obtain β estimates. Conversely, we suggest to minimize the modified distance function $\Delta_i(\alpha)y_i - \mu_i$ for the estimation of β.

(b) Performance of the MUWGEE approach

We now examine the performance of the modified distance function based MUWGEE approach (2.40). We also consider two more versions of this MUWGEE approach. First, we use a naive version of this approach by avoiding $\Sigma^*_{i(comp)}(\alpha^*)$ in (2.40). This leads (2.40) to be a modified unconditional weighted method of moments (MUWMM) estimating equation, given by

$$\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta}[\Delta_i(\alpha)y_i - \mu_i] = 0. \tag{2.42}$$

Secondly, we consider $C_i^*(\alpha^*)$ in $\Sigma^*_{i(comp)}(\alpha^*)$ in (2.40) as an identity matrix. This produces the 'working' independence (I) assumption based MUWGEE given by

$$\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta}[A_i]^{-1}[\Delta_i(\alpha)y_i - \mu_i] = \sum_{i=1}^{K} x_i'[\Delta_i(\alpha)y_i - \mu_i] = 0, \tag{2.43}$$

which, for convenience, we refer to as the MUWGEE(I). Note that this MUWGEE(I) is also a MM estimating equation where $x_i^{'}[\Delta_i(\alpha)y_i]$ is equated with to its unconditional expectation, namely $\sum_{i=1}^{K} x_i^{'}\mu_i$, to solve for β.

The simulated regression estimates based on the MUWMM (2.42) along with some what improved results produced by the MUWGEE (2.40) are shown in columns 3 and 4, and 5 and 6, respectively. For example, when $\rho = 0.6$, the MUWMM (2.42) yielded the estimates for $\beta_1 = \beta_2 = 0$ as $\hat{\beta}_1 = 0.104$ and $\hat{\beta}_2 = 0.370$, whereas the MUWGEE (2.40) produced slightly better estimates as $\hat{\beta}_1 = -0.070$ and $\hat{\beta}_2 = -0.216$. Note however that the MUWGEE approach still produces biased estimates. The standard errors produced by MUWGEE (2.40) are also relatively smaller yielding smaller SMSE, as compared to that of the MUWMM approach. Note that when these results, specially the estimates under the MUWGEE shown in columns 5 and 6 in Table 2.2 are compared with the corresponding UWGEE (2.38) [Robins et al. (1995)] estimates in columns 5 and 6 of Table 2.1, the modification, i.e., the use of $\Delta_i(\alpha)y_i - \mu_i$ instead of $\Delta_i(\alpha)(y_i - \mu_i)$ appears to reduce the bias to a large extent. As far as the performance of MUWGEE(I) approach is concerned, it is clear from Table 2.2 that MUWGEE approach produces much better estimates than this MUWGEE(I) approach.

(c) Performance of the CWGQL approach

Note that it was demonstrated based on the results of Table 2.1 and 2.2 that the modified unconditional distance function based GEE approach (namely MUWGEE (2.40)) performs better than the existing unconditional distance function based approach (UWGEE) of Robins et al. (1995). We now consider the proposed conditional weighted GQL estimating equation (2.34) and examine its performance to that of the MUWGEE approach. The simulation results based on the proposed simpler CWGQL approach are given in Table 2.3. In the same table, we also produce the estimates obtained by solving a conditional weighted MM (CWMM) estimating equation given

Table 2.2: Simulated mean (SM), simulated standard error (SSE) and simulated mean squared error (SMSE) for the MUWMM, MUWGEE(T) and MUWGEE(I) estimates with $\beta_1 = \beta_2 = 0.0$, $\alpha = 4$; under AR(1) longitudinal correlation structure for binary data with selected known values of ρ, based on 1000 simulations

		MUWMM		MUWGEE(T)		MUWGEE(I)	
ρ	Statistic	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
0.0	SM	-0.005	0.025	-0.005	0.025	-0.005	0.025
	SSE	0.409	0.660	0.408	0.661	0.408	0.661
	SMSE	0.167	0.436	0.167	0.437	0.167	0.437
0.2	SM	0.059	0.009	0.007	-0.132	0.059	0.010
	SSE	0.405	0.658	0.408	0.655	0.405	0.661
	SMSE	0.168	0.433	0.166	0.446	0.168	0.437
0.4	SM	0.084	0.099	-0.037	-0.225	0.083	0.099
	SSE	0.415	0.678	0.404	0.644	0.415	0.679
	SMSE	0.179	0.470	0.165	0.466	0.179	0.471
0.6	SM	0.104	0.370	-0.070	-0.216	0.103	0.372
	SSE	0.406	0.665	0.362	0.578	0.406	0.668
	SMSE	0.176	0.579	0.136	0.381	0.176	0.584
0.8	SM	0.227	0.734	-0.050	-0.169	0.225	0.736
	SSE	0.397	0.642	0.290	0.423	0.396	0.648
	SMSE	0.209	0.950	0.087	0.208	0.207	0.962

by

$$\sum_{i=1}^{K} x_i^{'}(\tilde{y}_i - \lambda_i) = 0. \tag{2.44}$$

Note that this CWMM estimating equation is obtained from the CWGQL approach by using $\rho = 0$ as well as $w_{it}(\alpha) = 1$ for all $i = 1, \cdots, K$ and $t = 1, \cdots, T_i$. This equation is comparable with the MUWGEE(I) given in (2.43), the later being obtained as a special case of the MUWGEE given in (2.40), whereas the CWMM is a similar special case of the CWGQL approach.

When the simulation results in Table 2.3 are compared with those of Table 2.2, it is clear that the proposed CWGQL approach produces almost unbiased estimates for both regression parameters, whereas the MUWGEE approach (see Table 2.2) produced the estimates with much larger biases as well as larger SMSEs. For example, when $\rho = 0.6$, the CWGQL produces MSEs as 0.114 and 0.312 for β_1 and β_2, whereas MUWGEE(T) (true correlation structure of y_{it} based MUWGEE) yielded larger MSEs 0.136 and 0.381 respectively. The existing UWGEE approach (see Table 2.1) performs much worse when compared to the CWGQL approach.

As far as the performance of CWMM approach (2.44) is concerned, this approach appears to be highly competitive to the proposed CWGQL approach, as compared to the other approaches discussed above.

2.6.3 Performance of the proposed CWGQL approach for unknown ρ

It is clear from the last section that the CWGQL approach produces the regression estimates with smaller MSE as compared to all other approaches including the existing UWGEE approach. This was however demonstrated for known longitudinal correlation ρ.

As ρ is unknown in practice, in this section we conduct another simulation study by estimating ρ by using

1. the unweighted (for missingness) variables based MM estimating equation (2.36)

Table 2.3: Simulated mean (SM), simulated standard error (SSE) and simulated mean squared error (SMSE) for the CWGQL and CWMM estimates with $\beta_1 = \beta_2 = 0.0$, $\alpha = 4$; under AR(1) longitudinal correlation structure for binary data with selected known values of ρ, based on 1000 simulations

		CWGQL		CWMM	
ρ	Statistic	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
0.0	SM	-0.011	0.024	-0.005	0.025
	SSE	0.387	0.620	0.408	0.661
	SMSE	0.150	0.385	0.167	0.437
0.2	SM	0.025	-0.018	0.034	0.012
	SSE	0.386	0.616	0.408	0.677
	SMSE	0.149	0.380	0.168	0.458
0.4	SM	0.008	-0.014	0.018	-0.006
	SSE	0.377	0.620	0.412	0.692
	SMSE	0.143	0.385	0.170	0.478
0.6	SM	-0.004	0.006	-0.016	0.033
	SSE	0.338	0.558	0.392	0.666
	SMSE	0.114	0.312	0.154	0.445
0.8	SM	0.005	0.015	0.010	-0.007
	SSE	0.270	0.447	0.366	0.564
	SMSE	0.073	0.200	0.134	0.318

2. the weighted variables based MM (WMM) estimating equation (2.37),

and then using the estimate in the CWGQL estimating equation in (2.34) for the estimation of β. The SM, SSE and SMSE for the estimates of β_1 and β_2 under this CWGQL approach with ρ estimated by MM or WMM approach are given in Table 2.4 for missingness indicator $\alpha = 4$, and in Table 2.5 for $\alpha = 1$. The SM, SSE and SMSE for the estimate of ρ are also shown in the same tables. The SMSEs for the estimates of ρ appear to be smaller when small values of ρ are estimated using the WMM approach, whereas for large ρ such as $\rho = 0.4$, 0.6 and 0.8, this approach produces estimates with larger MSE as compared to the MM approach. Thus, in general for correlated missing data, MM approach appears to be better than the WMM approach in estimating ρ.

With regard to the estimation of β_1 and β_2 for unknown ρ, the CWGQL estimates of β_1 and β_2 appear to have smaller MSE for both small and large ρ, when ρ is estimated by the MM approach. When the CWGQL estimates for β_1 and β_2 computed based on $\hat{\rho}_{MM}$, are compared with the same CWGQL estimates for known ρ, the estimates are found to be almost the same with slightly smaller MSE for the unknown ρ case. Note that the results from Table 2.4 and 2.5 show that the proposed CWGQL approach works quite well in estimating β_1 and β_2 even when ρ is unknown and estimated by the MM. For example, when $\alpha = 4$ and $\rho = 0.8$ is estimated by the MM approach, the results in Table 2.4 show that the CWGQL approach produces $\hat{\beta}_1 = -0.009$ and $\hat{\beta}_2 = -0.052$ which are very close to the corresponding true values $\beta_1 = 0$ and $\beta_2 = 0$, respectively. Since the standard errors of these estimates are also found to be small, the CWGQL approach appears to perform very well in general in obtaining unbiased and hence consistent estimates for the regression effects.

Table 2.4: Simulated mean (SM), simulated standard error (SSE) and simulated mean squared error (SMSE) for the CWGQL estimates of β under binary data when AR(1) correlation parameter ρ is estimated by a selected method of moments, with $\beta_1 = \beta_2 = 0.0$, $\alpha = 4$, based on 1000 simulations

ρ	Statistic	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\rho}_{MM}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\rho}_{WMM}$
0.0	SM	-0.058	-0.058	0.119	-0.009	0.019	-0.007
	SSE	0.388	0.609	0.074	0.389	0.624	0.075
	SMSE	0.154	0.374	0.020	0.151	0.389	0.006
0.2	SM	-0.014	-0.115	0.287	0.044	0.017	0.153
	SSE	0.383	0.597	0.069	0.391	0.631	0.076
	SMSE	0.147	0.369	0.012	0.155	0.398	0.008
0.4	SM	-0.021	-0.098	0.456	0.043	0.095	0.322
	SSE	0.372	0.602	0.062	0.389	0.671	0.080
	SMSE	0.139	0.372	0.007	0.153	0.459	0.013
0.6	SM	-0.020	-0.056	0.626	0.041	0.194	0.507
	SSE	0.329	0.533	0.056	0.358	0.646	0.080
	SMSE	0.109	0.288	0.004	0.130	0.455	0.015
0.8	SM	-0.009	-0.052	0.797	0.053	0.144	0.721
	SSE	0.261	0.351	0.042	0.302	0.478	0.070
	SMSE	0.068	0.126	0.002	0.094	0.249	0.011

Table 2.5: Simulated mean (SM), simulated standard error (SSE) and simulated mean squared error (SMSE) for the CWGQL estimates of β under binary data when AR(1) correlation parameter ρ is estimated by a selected method of moments, with $\beta_1 = \beta_2 = 0.0$, $\alpha = 1$, based on 1000 simulations

ρ	Statistic	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\rho}_{MM}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\rho}_{WMM}$
0.0	SM	-0.021	0.054	0.098	-0.011	0.009	-0.004
	SSE	0.388	0.741	0.076	0.387	0.742	0.075
	SMSE	0.151	0.552	0.015	0.150	0.551	0.006
0.2	SM	0.002	0.048	0.237	0.012	0.008	0.136
	SSE	0.404	0.770	0.071	0.408	0.777	0.076
	SMSE	0.163	0.595	0.006	0.167	0.603	0.010
0.4	SM	-0.006	-0.019	0.374	0.000	-0.042	0.286
	SSE	0.399	0.744	0.068	0.405	0.771	0.083
	SMSE	0.159	0.553	0.005	0.164	0.596	0.020
0.6	SM	0.012	-0.091	0.514	0.016	-0.108	0.448
	SSE	0.390	0.738	0.062	0.402	0.787	0.080
	SMSE	0.152	0.553	0.011	0.162	0.631	0.029
0.8	SM	-0.050	-0.098	0.649	-0.045	-0.103	0.618
	SSE	0.369	0.658	0.052	0.374	0.693	0.067
	SMSE	0.138	0.442	0.025	0.142	0.490	0.038

2.6.4 Estimation of α: The effects of past observations on response indication

Recall from (2.7) that for $q = 1$, $\alpha \equiv \alpha_1$ represents the dependence of the response indicator at the current time on the past response. So far, we have assumed that this parameter (α) is however known. In practice it is likely that α is unknown. In this section, we provide a 'working' likelihood approach to estimate α, by treating the available R_{it}'s as fixed responses and the past responses $\{y_{i,t-1}\}$ as the known covariates. But, we first give an interpretation of this parameter (α) .

Interpretation of α:

It is clear from (2.7) that if α is large, then the response probability will also be large. In order to see how a change in α value changes the proportion of non-missing, for a given α, we simply compute the proportion of non-missing $\hat{p}_{NM}^{(s)}$, in the s-th ($s = 1, \cdots, 1000$) simulation, by using the formula

$$\hat{p}_{NM}^{(s)} = \frac{\sum_{i=1}^{K} T_i^{(s)}}{\sum_{i=1}^{K} S_i^{(s)}}, \tag{2.45}$$

where $T_i^{(s)}$ is the total number of times that the i-th individual responded in the s-th simulation and $S_i^{(s)}$ is either $T_i^{(s)}$ when $T_i^{(s)} = T$, or $T_i^{(s)} + 1$ if $T_i^{(s)} < T$. Note that $S_i^{(s)} = T_i^{(s)} = T$ means that the i-th individual responded in all follow ups, whereas $S_i^{(s)} = T_i^{(s)} + 1$ indicates the number of attempts to have $T_i^{(s)}$ response from the i-th individual. We next take the average of $\hat{p}_{NM}^{(s)}$ over all simulations and compute the estimate of proportion of non-missing (p_{NM}) as

$$\hat{p}_{NM} = \frac{\sum_{s=1}^{1000} \hat{p}_{NM}^{(s)}}{1000}. \tag{2.46}$$

Note that we have computed $\hat{p}_{NM}$ (2.46) under the simulation study reported in Tables 2.4 and 2.5 and found that $\hat{p}_{NM} = 0.907$ when $\alpha = 4$, and $\hat{p}_{NM} = 0.858$ when $\alpha = 1$. Thus $\alpha = 4$ based on the simulation designs for Table 2.4 or 2.5 indicates 9% missing, whereas $\alpha = 1$ based on the same simulation design exhibits 14% missing

Table 2.6: Proportion of non-missing (p_{NM}) for various values of α under the same simulation design for binary data

α	1	2	3	4	5
$\hat{p}_{NM}$	0.858	0.887	0.900	0.907	0.908

Table 2.7: Simulated mean (SM) and simulated median (SMed) of the likelihood estimates of α based on (2.47) for binary data

α	1	2	3	4
$\hat{\alpha}$ based on SM	1.028	2.142	4.343	9.139
$\hat{\alpha}$ based on SMed	1.003	2.017	3.127	3.984

responses. For the same simulation design parameters as for Table 2.4 and 2.5, we have also computed $\hat{p}_{NM}$ by (2.46) for some other values of α. The values of $\hat{p}_{NM}$ for the corresponding α values are shown in Table 2.6.

Estimation of α:

We now return to the estimation of α. For this purpose, similar to Robins et al. [1995, eqn. (9), p.109], we first construct a 'working' likelihood function by treating the available R_{it}'s as fixed binary responses and the past data $\{y_{i,t-1}\}$ as the known covariates. The log-likelihood is given by

$$logL = \sum_{i=1}^{K} \sum_{t=1}^{min[T_i+1,T]} \{R_{it} log[g_{it}(\alpha \mid y_{i,t-1})] + (1 - R_{it}) log[1 - g_{it}(\alpha \mid y_{i,t-1})]\}, \quad (2.47)$$

where $g_{it}(\alpha \mid y_{i,t-1})$ is given by (2.7), with y_{it} generated from the longitudinal binary model subject to MAR, i.e. y_{it} is generated as $y_{it} \sim bin(\mu^*_{it})$, with

$$\mu^*_{it} = \lambda_{it|t-1}(\beta, \rho) w_{it}(\alpha). \quad (2.48)$$

We then solve $(\partial logL/\partial\alpha) = 0$ for the likelihood estimate of α. For the same simulation designs as considered for the results in Tables 2.1 through 2.5, we use selected values of α, namely $\alpha = 1, 2, 3$, and 4, and obtained the so-called likelihood estimate of α as in Table 2.7. Note that the likelihood (2.47) based simulated estimates of α appeared to be skewed, specially when α is large. Thus the simulated median (SMed) appears to reflect well the true value of the α parameter.

Chapter 3

Incomplete Longitudinal Models for Count Data

In Chapter 2, we analyzed regression models for the incomplete longitudinal binary data. As mentioned in Chapter 1, in practice there are many situations where one may have to deal with longitudinal count data instead of binary data. For example, in a bio-medical longitudinal study, one may be interested to analyze the number of yearly visits by selected individuals, to their physicians over a period of 4 or 5 years. Here it may be of importance to know the effects of associated covariates such as gender and education level on the number of visits that are correlated. In this problem, it is likely that a few responses of some of the individuals may be missing at random. Thus, in practice incomplete longitudinal count data analysis may also be of interest. Note however that unlike the binary model, there is no adequate discussion on the incomplete longitudinal count data models. In this chapter, we study such count data model for the incomplete situation.

Note that for modelling correlations of the longitudinal count data it may be appropriate to consider an exponentially decaying correlation structure based AR(1) type model. This type of model for the complete repeated count data is given by

$$y_{it} = \rho * y_{i,t-1} + d_{it}, \ t = 2, \cdots, T \tag{3.1}$$

[see eqn. (1.21)] where for given $y_{i,t-1}$, $\rho * y_{i,t-1}$ is computed by using the binomial thinning operation given by

$$\rho * y_{i,t-1} = \sum_{j=1}^{y_{i,t-1}} b_j(\rho)$$

with $P[b_j(\rho) = 1] = \rho$ and $P[b_j(\rho) = 0] = 1 - \rho$. It may be shown that the mean of y_{it} conditional on $y_{i,t-1}$ is given by

$$E(Y_{it} \mid y_{i,t-1}) = E\left[\left\{\sum_{j=1}^{y_{i,t-1}} b_j(\rho) + d_{it}\right\} \mid y_{i,t-1}\right]. \tag{3.2}$$

Note that for given $y_{i,t-1}$, $\sum_{j=1}^{y_{i,t-1}} b_j(\rho)$ follows the binomial distribution with size $y_{i,t-1}$ and probability of success ρ. Consequently, $E\left[\sum_{j=1}^{y_{i,t-1}} b_j(\rho) \mid y_{i,t-1}\right] = y_{i,t-1}\rho$. Since d_{it} follows the Poisson distribution with mean parameter $\mu_{it} - \rho\mu_{i,t-1}$ as assumed in (1.21), it is clear that $E(d_{it}) = \mu_{it} - \rho\mu_{i,t-1}$. Furthermore, as $y_{i,t-1}$ and d_{it} are assumed to be independent, it then follows from (3.2) that

$$\begin{aligned} E(Y_{it} \mid y_{i,t-1}) &= y_{i,t-1}\rho + (\mu_{it} - \rho\mu_{i,t-1}) \\ &= \mu_{it} + \rho(y_{i,t-1} - \mu_{i,t-1}) \\ &= \lambda_{it|t-1}(\beta, \rho). \end{aligned} \tag{3.3}$$

It is interesting to point out that the conditional mean, i.e. $E(Y_{it} \mid y_{i,t-1})$ under both binary model (1.26) and Poisson model (3.1) have the same form except that $\mu_{it} = exp(x'_{it}\beta)/[1 + exp(x'_{it}\beta)]$ in the binary case and $\mu_{it} = exp(x'_{it}\beta)$ in the Poisson case. There is however a difference in the ranges for the ρ parameter. In the binary case, ρ lies in the range as shown in (1.32), whereas in the count data case, ρ satisfy the range restriction

$$0 < \rho < min_{(i,t)}[1, \mu_{it}/\mu_{i,t-1}].$$

Note that by taking further expectation over (3.3) with respect to $y_{i,t-1}$, one obtains the unconditional mean of y_{it} as μ_{it} given in (1.22). To compute the unconditional variance, we use

$$V(Y_{it}) = E_{Y_{i,t-1}} V(Y_{it} \mid Y_{i,t-1}) + V_{Y_{i,t-1}} E(Y_{it} \mid Y_{i,t-1})$$

$$
\begin{aligned}
&= E_{Y_{i,t-1}}\left[Y_{i,t-1}\rho(1-\rho)+\mu_{it}-\rho\mu_{i,t-1}\right]+V_{Y_{i,t-1}}\left[\mu_{it}+\rho(y_{i,t-1}-\mu_{i,t-1})\right] \\
&= \rho(1-\rho)E_{Y_{i,t-1}}(Y_{i,t-1})+\mu_{it}-\rho\mu_{i,t-1}+\rho^2 V_{Y_{i,t-1}}(Y_{i,t-1}) \\
&= \rho\mu_{i,t-1}-\rho^2\mu_{i,t-1}+\mu_{it}-\rho\mu_{i,t-1}+\rho^2\mu_{i,t-1} \\
&= \mu_{it}. \qquad (3.4)
\end{aligned}
$$

Next by using the dynamic model (3.1), one may show that

$$
\begin{aligned}
E(Y_{it}Y_{i,t-1}) &= E_{Y_{i,t-1}}\left[Y_{it}Y_{i,t-1} \mid Y_{i,t-1}\right] \\
&= E_{Y_{i,t-1}}\left[Y_{i,t-1}E(Y_{it} \mid Y_{i,t-1})\right] \\
&= E_{Y_{i,t-1}}\left[Y_{i,t-1}\{\mu_{it}+\rho(Y_{i,t-1}-\mu_{i,t-1})\}\right] \\
&= \mu_{it}\mu_{i,t-1}+\rho(\mu_{i,t-1}-\mu_{i,t-1}^2)-\rho\mu_{i,t-1}^2 \\
&= \rho\mu_{i,t-1}+\mu_{it}\mu_{i,t-1}. \qquad (3.5)
\end{aligned}
$$

By similar calculation, it may be shown that for $t < t^{'}$

$$E(Y_{it}Y_{it'}) = \rho^{(t^{'}-t)}\mu_{it}+\mu_{it}\mu_{it'}. \qquad (3.6)$$

Hence, it follows that

$$cov(Y_{it}Y_{it'}) = \rho^{(t^{'}-t)}\mu_{it}, \qquad (3.7)$$

implying that the correlation between y_{it} and $y_{it'}$ is given by

$$corr(Y_{it},Y_{it'}) = \rho^{(t^{'}-t)}\left[\frac{\mu_{it}}{\mu_{it'}}\right]^{1/2} \text{ for } t < t^{'},$$

as pointed out in (1.22).

3.1 Inferences in Conditional Probability Models for Incomplete Longitudinal Count Data

When the repeated count data are considered to be MAR (a) the weighted variable given by $\tilde{y}_{it} = \delta_{it}(\alpha)y_{it}$, (b) the unconditional distance function given by $\delta_{it}(\alpha)(y_{it} - \mu_{it})$ (2.17), and (c) the modified unconditional distance function given by $\delta_{it}(\alpha)y_{it} - \mu_{it}$ (2.15) remain the same as those under the binary data case given in (2.23), (2.17) and (2.15), respectively. Consequently, the forms of the estimating equations discussed under the incomplete binary data model, namely UWGEE (2.38), MUWGEE (2.40), CWGQL (2.34) and CWMM (2.44) remain the same for the incomplete count data model. The difference lies in the specific formulas for the derivatives involved in these forms between the binary and the count data cases. An additional difference with regard to the formulation of the weight matrix (inverse of the covariance matrix) arises only for the CWGQL approach. More specifically, the computation of the weight matrix involved in the CWGQL (2.34) would be different under the count data model as compared to the binary model. This is because the count data model (3.1) and the conditional linear binary dynamic (CLBD) model (1.26) provide different variances and covariances for the weighted responses. We now provide the construction of such weight matrix under the count data model (3.1), whereas similar construction for the binary case was done in (2.25) and (2.26).

3.1.1 Construction of the CWGQL estimation approach for the count data model

The form of the CWGQL estimating equation will remain the same as given in the incomplete binary case by (2.34). However, $\tilde{\Sigma}_{iw}$ in (2.34) will be different because of the difference in the longitudinal correlation models between the binary and the count data cases. For the present count data case, the conditional variance of the weighted response variable $\tilde{y}_{it}$ may be computed as

$$\begin{aligned}
var(\tilde{Y}_{it} \mid H_{i,t-1}) &= E(\tilde{Y}_{it}^2 \mid H_{i,t-1}) - \lambda^2_{it|t-1}(\beta,\rho) \\
&= E\left[\left\{\frac{R_{it}Y_{it}}{w_{it}(\alpha)}\right\}^2 \mid H_{i,t-1}\right] - \lambda^2_{it|t-1}(\beta,\rho) \\
&= \frac{1}{w_{it}^2(\alpha)} E\left(R_{it}Y_{it}^2 \mid H_{i,t-1}\right) - \lambda^2_{it|t-1}(\beta,\rho) \\
&= \frac{1}{w_{it}^2(\alpha)} E(R_{it} \mid H_{i,t-1}) E(Y_{it}^2 \mid H_{i,t-1}) - \lambda^2_{it|t-1}(\beta,\rho) \\
&= \frac{1}{w_{it}^2(\alpha)} w_{it}(\alpha)\left[V(Y_{it} \mid H_{i,t-1}) + E^2(Y_{it} \mid H_{i,t-1})\right] - \lambda^2_{it|t-1}(\beta,\rho) \\
&= \frac{1}{w_{it}(\alpha)}\left[y_{i,t-1}\rho(1-\rho) + \mu_{it} - \rho\mu_{i,t-1} + \lambda^2_{it|t-1}(\beta,\rho)\right] - \lambda^2_{it|t-1}(\beta,\rho) \\
&= \frac{1}{w_{it}(\alpha)}[\lambda_{it|t-1}(\beta,\rho) - \rho^2 y_{i,t-1} + \lambda^2_{it|t-1}(\beta,\rho)] - \lambda^2_{it|t-1}(\beta,\rho) \\
&= \frac{1}{w_{it}(\alpha)}[\lambda_{it|t-1}(\beta,\rho)\{1 + \lambda_{it|t-1}(\beta,\rho)\} \\
&\qquad\qquad -\rho^2 y_{i,t-1}] - \lambda^2_{it|t-1}(\beta,\rho). \qquad (3.8)
\end{aligned}$$

Note that the main reason for this formula in (3.8) to be different than the corresponding formula (2.25) for the binary case is that unlike in the count data case, $E\left[(R_{it}Y_{it})^2\right] = E(R_{it}Y_{it})$ in the binary case. Further note that the conditional covariances between the weighted count responses at two different time points are still zero following the same argument as in the binary data case given by (2.26), i.e., for $l = max(t,t'),\ t \neq t',\ t,t' = 1,\cdots,T$

$$cov(\tilde{Y}_{it}, \tilde{Y}_{it'} \mid H_{i,l-1}) = 0. \qquad (3.9)$$

Thus $\tilde{\Sigma}_{iw} = (\tilde{\sigma}_{iut})$ for the count data is constructed by using (3.8) and (3.9), whereas in the binary data case this was constructed based on (2.25) and (2.26). Consequently, similar to (2.34), the CWGQL estimating equation is given by

$$\sum_{i=1}^{K} \frac{\partial \lambda_i'}{\partial \beta} [\tilde{\Sigma}_{iw}]^{-1} [\tilde{y}_i - \lambda_i] = 0, \tag{3.10}$$

but, also unlike in (2.34), $\left[\partial \lambda_{i,t|t-1}(\beta, \rho)/\partial \beta_s\right]$ in the present case has the formula

$$\frac{\partial \lambda_{i,t|t-1}(\beta,\rho)}{\partial \beta_s} = \begin{cases} x_{i1s}\mu_{i1}, & \text{for } t = 1 \\ x_{its}\mu_{it} - \rho x_{i,t-1,s}\mu_{i,t-1}, & \text{for } t > 1. \end{cases} \tag{3.11}$$

Note that the CWMM equation for the count data model has the same formula as in (2.44) for the binary data, except that $\mu_{it} = exp(x_{it}'\beta)/[1 + exp(x_{it}'\beta)]$ in the binary case and $\mu_{it} = exp(x_{it}'\beta)$ in the Poisson case.

3.2 Simulation Study for the Incomplete Longitudinal Count Data

To examine the relative performance of the proposed as well as existing estimating equation approaches, we now conduct a Monte Carlo study based on 1000 simulations. We consider $K = 100$ individuals and for $T = 4$, generate $\sum_{i=1}^{K} T_i$ responses based on MAR mechanism following the same steps given in Section 2.6.1 with $q = 1$, $\alpha = 3$ and/or 1, and AR(1) longitudinal correlation structure (3.3) for the count data with $\rho = 0.2$, 0.4, 0.6 and 0.8. As far as the covariates are concerned, we consider $p = 2$ with x_{it1} and x_{it2} defined as

$$x_{it1} = \begin{cases} 0 & \text{for } i = 1, \cdots, \frac{K}{4};\ t = 1, 2 \\ \frac{1}{2} & \text{for } i = 1, \cdots, \frac{K}{4};\ t = 3, 4 \\ -\frac{1}{2} & \text{for } i = \frac{K}{4} + 1, \cdots, \frac{3K}{4};\ t = 1 \\ 0 & \text{for } i = \frac{K}{4} + 1, \cdots, \frac{3K}{4};\ t = 2, 3 \\ \frac{1}{2} & \text{for } i = \frac{K}{4} + 1, \cdots, \frac{3K}{4};\ t = 4 \\ \frac{t}{2T} & \text{for } i = \frac{3K}{4} + 1, \cdots, K;\ t = 1, \cdots, 4 \end{cases}$$

and

$$x_{it2} = \begin{cases} \frac{t}{2T} & \text{for } i = 1, \cdots, \frac{K}{2};\ t = 1, \cdots, 4 \\ 0 & \text{for } i = \frac{K}{2}+1, \cdots, K;\ t = 1, 2 \\ \frac{1}{2} & \text{for } i = \frac{K}{2}+1, \cdots, K;\ t = 3, 4 \end{cases}$$

respectively, and use $\beta_1 = \beta_2 = 0.5$.

3.2.1 Estimation of β for known ρ

To examine the performance of the various approaches under the incomplete count data model in estimating the regression effects, we first assume that the longitudinal correlation parameter ρ is known. Next, we assume that the dependence parameter α of the non-response model (2.7) is also known as $\alpha = 3$. Note that $\alpha = 3$ represents 12% missing responses, whereas $\alpha = 1$ represents 16% missing responses under the count data model.

(a) Performance of the UWGEE approach

The SMs, SSEs and SMSEs of the estimates of β based on 1000 simulations for the UWGEE (2.38) approach under the incomplete count data model are reported in Table 3.1. Similar to the binary data case, here we also have considered the true AR(1) as well as several 'working' correlation structures namely, independence, MA(1) and EQC based UWGEE (2.38). It is clear that UWGEE approach of Robins et al. (1995) produces estimates of β with large bias irrespective of the selected values of ρ and selected correlation structure. This is not surprising as we have also found similar performance of this approach under the incomplete binary model discussed in Chapter 2. Recall that we argued in Chapter 2 that the reasons of this poor performance of UWGEE approach are (i) inappropriate construction of the unconditional distance function, and (ii) the use of correlation structure of the unweighted random variable y_{it} instead of the use of appropriate correlation structure of the weighted random variable $\tilde{y}_{it} = \delta_{it}(\alpha)y_{it}$. Note that while we compute the UWGEE (2.38) based estimates under

Table 3.1: Simulated mean (SM), simulated standard error (SSE) and simulated mean squared error (SMSE) for the UWGEE based estimates with $\beta_1 = \beta_2 = 0.5$, $\alpha = 3$ and selected known values of ρ under true AR(1) longitudinal correlation structure for count data, based on 1000 simulations

		True		Working correlation structures					
		AR(1)		(I)		(EQC)		(MA(1))	
ρ	Statistic	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
0.2	SM	0.349	-0.025	0.383	-0.013	0.317	-0.043	0.348	-0.019
	SSE	0.212	0.243	0.222	0.251	0.207	0.242	0.211	0.243
	SMSE	0.068	0.335	0.063	0.326	0.076	0.354	0.068	0.329
0.4	SM	0.300	-0.024	0.374	0.005	0.252	-0.049	0.295	0.001
	SSE	0.207	0.257	0.235	0.287	0.208	0.265	0.210	0.260
	SMSE	0.083	0.341	0.071	0.327	0.104	0.371	0.086	0.316
0.6	SM	0.250	-0.003	0.387	0.051	0.198	-0.024	–	–
	SSE	0.212	0.262	0.250	0.314	0.214	0.280	–	–
	SMSE	0.107	0.322	0.076	0.300	0.137	0.353	–	–
0.8	SM	0.203	0.045	0.402	0.132	0.174	0.024	–	–
	SSE	0.207	0.257	0.266	0.349	0.227	0.306	–	–
	SMSE	0.131	0.273	0.077	0.257	0.158	0.320	–	–

incomplete count data model, the construction of the derivative matrix, $\partial \mu_i'/\partial \beta$ is different than that of the binary data case. To be specific, the derivative of the t-the element of μ_i with respect to the s-th ($s = 1, \cdots, p$) element of β has the formula

$$\frac{\partial \mu_{it}}{\partial \beta_s} = x_{its}\mu_{it}, \tag{3.12}$$

which is different than (2.39) under the binary data model.

(b) Performance of the MUWGEE approach

We now examine the performance of the MUWGEE approach (2.40) under the count data model. Recall that this MUWGEE approach is a modification of the existing UWGEE (2.38) approach of Robins et al. (1995), where the modification in the unconditional distance function is suggested for the use of $[\delta_{it}(\alpha)y_{it} - \mu_{it}]$ instead of

Table 3.2: Simulated mean (SM), simulated standard error (SSE) and simulated mean squared error (SMSE) for the MUWGEE based estimates with $\beta_1 = \beta_2 = 0.5$, $\alpha = 3$ and selected known values of ρ, under true AR(1) longitudinal correlation structure for count data, based on 1000 simulations

		MUWGEE(T)		MUWGEE(I)	
ρ	Statistic	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
0.2	SM	0.489	0.452	0.489	0.496
	SSE	0.216	0.227	0.217	0.225
	SMSE	0.047	0.054	0.047	0.051
0.4	SM	0.478	0.420	0.486	0.519
	SSE	0.219	0.250	0.228	0.257
	SMSE	0.048	0.069	0.052	0.066
0.6	SM	0.467	0.403	0.500	0.574
	SSE	0.224	0.264	0.243	0.279
	SMSE	0.051	0.079	0.059	0.083
0.8	SM	0.458	0.408	0.523	0.668
	SSE	0.216	0.273	0.248	0.308
	SMSE	0.049	0.083	0.062	0.123

$\delta_{it}(\alpha)(y_{it} - \mu_{it})$. Note that this modification appeared to improve the regression estimates to some extent over the existing UWGEE approach (2.38) under the incomplete binary model (see Section 2.6.2). Under the incomplete count data model, this modification works quite well in obtaining unbiased regression estimates. This is clear from the results reported in Table 3.2 where the simulation results for MUWGEE approach are reported under the count data model. To be specific, for known $\rho = 0.6$, the true AR(1) structure based MUWGEE approach [MUWGEE(T)] yielded the estimates of β as $\hat{\beta} = (0.467, 0.403)'$ which is close to the true values $\beta = (0.5, 0.5)'$ indicating significant improvement in bias reduction, whereas from Table 3.1, the existing UWGEE approach yielded $\hat{\beta} = (0.250, -0.003)'$ which is far from the true values. The independece assumption based MUWGEE approach [MUWGEE(I)] also improves the estimates over UWGEE approach, but trails behind MUWGEE(T) approach.

We now discuss the proposed CWGQL approach which is expected to perform the same or better than the MUWGEE approach.

(c) Performance of the CWGQL approach

Recall that the MUWGEE approach (2.40) was constructed by correcting only the unconditional distance function of the existing UWGEE approach of Robins et al. (1995), but retaining the same 'working' covariance as in the UWGEE approach. Even though the biases were found to be small under the MUWGEE approach (2.40), the use of proper unconditional correlation matrix of the weighted variable $\tilde{y}_{it} = \delta_{it}(\alpha)y_{it}$, as opposed to the 'working' correlation matrix of y_{it}, may further reduce the bias as well as the standard errors as compared to the MUWGEE approach. But the construction and the computation of such unconditional correlation matrix under the count data model is very complicated. We therefore proposed the CWGQL approach (2.34) which is constructed based on the conditional distance function $[\delta_{it}y_{it} - \lambda_{it|t-1}(\beta, \rho)]$ and also accommodates both missing mechanism and longitudinal correlation structure for the count data through the weighted variables. As discussed in Chapter 2, this CWGQL approach produced consistent and efficient estimates of the regression parameters under the incomplete binary model. We now examine the performance of the proposed CWGQL approach (2.34) under the present incomplete count data model.

The simulation results of CWGQL approach are reported in columns 3 and 4 of Table 3.3 under the incomplete count data model. In columns 5 and 6 of the same table, we have also reported the simulation results under the CWMM approach (2.44). It is clear from the table that the CWGQL approach produces almost unbiaesd regression estimates with smaller standard errors as compared to any other approaches considered in the thesis. To be specific, when the CWGQL estimates are compared with the corresponding MUWGEE approach based estimates which was found to be the best among others under incomplete count data model, we found that the CWGQL approach produces estimates with much less biases and standard errors. For example, when $\rho = 0.6$ the CWGQL estimates are obtained as $\hat{\beta} = (0.506, 0.490)'$

Table 3.3: Simulated mean (SM), simulated standard error (SSE) and simulated mean squared error (SMSE) for the CWGQL and CWMM estimates with $\beta_1 = \beta_2 = 0.5$, $\alpha = 3$; under AR(1) longitudinal correlation structure for the count data with selected known values of ρ, based on 1000 simulations

		CWGQL		CWMM	
ρ	Statistic	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
0.2	SM	0.501	0.491	0.500	0.490
	SSE	0.211	0.220	0.215	0.224
	SMSE	0.045	0.048	0.046	0.050
0.4	SM	0.508	0.483	0.506	0.486
	SSE	0.210	0.240	0.218	0.249
	SMSE	0.044	0.058	0.048	0.062
0.6	SM	0.506	0.490	0.509	0.483
	SSE	0.206	0.246	0.230	0.264
	SMSE	0.042	0.061	0.053	0.070
0.8	SM	0.508	0.497	0.506	0.490
	SSE	0.197	0.249	0.224	0.281
	SMSE	0.039	0.062	0.050	0.079

with SMSEs as $(0.042, 0.061)'$, whereas MUWGEE(T) produced regression estimates as $\hat{\beta} = (0.467, 0.403)'$ with SMSEs as $(0.051, 0.079)'$. Note that this result under the CWGQL approach is a large improvement in terms of bias and standard error over the MUWGEE approach or the existing UWGEE approach. As far as the performance of the CWMM estimates are concerned, this approach also performs well, but the CWGQL approach remains the best as compared to its all competitors.

3.2.2 Performance of the proposed CWGQL approach for unknown ρ

As ρ is unknown in practice, in this section we conduct another simulation study considering ρ to be unknown and estimated by a suitable MM estimator. Recall that in the binary data case, we have given two MM approaches, namely $\hat{\rho}_{MM}$ (2.36) and $\hat{\rho}_{WMM}$ (2.37) in estimating this correlation index parameter. Note that these formulas under the incomplete count data model remain the same except the means and variances are different under the count data case. To be specific, here $\mu_{it} = \sigma_{i,tt} = exp(x_{it}^{'}\beta)$.

In the simulation study, we first estimate ρ by $\hat{\rho}_{MM}$ or $\hat{\rho}_{WMM}$ and use that estimate to obtain the CWGQL estimate of β. The similation results under the CWGQL approach with estimated ρ by both $\hat{\rho}_{MM}$ and $\hat{\rho}_{WMM}$ are shown in Table 3.4 for $\alpha = 3$ and in Table 3.5 for $\alpha = 1$. Similar to binary data case, the WMM approach estimates ρ with smaller SMSEs as compared the MM approach for small correlations such as $\rho = 0.2$ and 0.4, but for large correlations MM approach produces ρ estimates with smaller SMSEs. With regard to the estimation of CWGQL approach, the performance of the estimation of β are almost the same irrespective of the use of MM or WMM estimates of ρ. For example, when $\alpha = 3$ and $\rho = 0.8$, the SMSEs of $\beta = (\beta_1, \beta_2)'$ were found to be $(0.038, 0.064)'$ when $\hat{\rho}_{MM}$ was used, whereas the use of $\hat{\rho}_{WMM}$ produced the CWGQL estimates with SMSEs as $(0.040, 0.065)'$. Therefore, the CWGQL approach produces consistent as well as efficient regression estimates under the incomplete count data model when ρ is estimated by either MM approach or WMM approach.

3.2.3 Estimation of α: The effects of past observations on response indication

Similar to the binary case, we first examine here the proportion of missing observations for a given value of α. To do this, unlike the binary case we compute the proportion of non-missing $\hat{p}_{NM}^{(s)}$, in the s-th ($s = 1, \cdots, 1000$) simulation, by using the formula

$$\hat{p}_{NM}^{(s)} = \frac{1}{2}\left[\frac{\sum_{i=1}^{K} T_i^{(s)}}{\sum_{i=1}^{K} S_i^{(s)}} + \frac{\sum_{i=1}^{K} T_i^{(s)}}{KT}\right], \qquad (3.13)$$

where $T_i^{(s)}$ and $S_i^{(s)}$ are as in (2.45) under the binary model. Note that (3.13) is slightly different than (2.45). Since the ratio of $\sum_{i=1}^{K} T_i^{(s)}$ (total number of observed

Table 3.4: Simulated mean (SM), simulated standard error (SSE) and simulated mean squared error (SMSE) for the CWGQL estimates of β for count data when AR(1) correlation parameter ρ is estimated by a selected method of moments, with $\beta_1 = \beta_2 = 0.5$, $\alpha = 3$, based on 1000 simulations

ρ	Statistic	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\rho}_{MM}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\rho}_{WMM}$
0.2	SM	0.501	0.477	0.299	0.501	0.493	0.171
	SSE	0.212	0.220	0.075	0.211	0.220	0.077
	SMSE	0.045	0.049	0.015	0.045	0.048	0.007
0.4	SM	0.506	0.475	0.468	0.509	0.485	0.351
	SSE	0.209	0.239	0.079	0.212	0.243	0.086
	SMSE	0.044	0.058	0.011	0.045	0.059	0.010
0.6	SM	0.505	0.484	0.631	0.511	0.492	0.540
	SSE	0.206	0.245	0.074	0.208	0.248	0.084
	SMSE	0.042	0.060	0.006	0.044	0.062	0.011
0.8	SM	0.510	0.497	0.768	0.519	0.501	0.709
	SSE	0.195	0.252	0.068	0.198	0.255	0.077
	SMSE	0.038	0.064	0.006	0.040	0.065	0.014

responses) to the total number for complete data case (KT) also indicates a non-missing proportion, in (3.13), we have used an average of this ratio and the former ratio ($\sum_{i=1}^{K} T_i^{(s)} / \sum_{i=1}^{K} S_i^{(s)}$) to understand the overall non-missing proportion in the count data case. The simulated average of $\hat{p}_{NM}^{(s)}$ based on 1000 simulations, i.e.,

$$\hat{p}_{NM} = \frac{\sum_{s=1}^{1000} \hat{p}_{NM}^{(s)}}{1000} \tag{3.14}$$

are reported in Table 3.6.

Estimation of α:

Note that when it is needed to estimate α for the count data case, we maximize the log-likelihood function (2.47) as in the binary case. However, it is important to recognize that the past responses involved in the log-likelihood function (2.47) are

Table 3.5: Simulated mean (SM), simulated standard error (SSE) and simulated mean squared error (SMSE) for the CWGQL estimates of β for count data when AR(1) correlation parameter ρ is estimated by a selected method of moments, with $\beta_1 = \beta_2 = 0.5$, $\alpha = 1$, based on 1000 simulations

ρ	Statistic	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\rho}_{MM}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\rho}_{WMM}$
0.2	SM	0.505	0.498	0.308	0.498	0.490	0.167
	SSE	0.215	0.235	0.080	0.216	0.237	0.080
	SMSE	0.046	0.055	0.018	0.047	0.056	0.007
0.4	SM	0.515	0.503	0.473	0.512	0.485	0.343
	SSE	0.221	0.254	0.086	0.225	0.261	0.093
	SMSE	0.049	0.064	0.013	0.051	0.068	0.012
0.6	SM	0.527	0.490	0.634	0.525	0.467	0.532
	SSE	0.217	0.264	0.086	0.221	0.271	0.095
	SMSE	0.048	0.070	0.009	0.049	0.075	0.014
0.8	SM	0.499	0.492	0.780	0.496	0.471	0.712
	SSE	0.205	0.268	0.080	0.208	0.277	0.088
	SMSE	0.042	0.072	0.007	0.043	0.078	0.016

Table 3.6: Proportion of non-missing (p_{NM}) for various values of α under the same simulation design for count data

α	1	2	3	4	5
$\hat{p}_{NM}$	0.837	0.868	0.880	0.883	0.885

Table 3.7: Simulated mean (SM) and simulated median (SMed) of the likelihood estimates of α based on (2.47) for count data

α	1	2	3	4
$\hat{\alpha}$ based on SM	1.036	2.225	4.579	6.375
$\hat{\alpha}$ based on SMed	1.010	2.025	2.894	3.453

now generated from a Poisson distribution with mean parameter μ^*_{it} given by

$$\mu^*_{it} = \lambda_{it|t-1}(\beta, \rho) w_{it}(\alpha) \tag{3.15}$$

which has the same form as in (2.48), but they are different as they contain count or binary data depending on the situation.

The simulated mean and median for the estimates of α are reported in Table 3.7. Note that similar to the binary data case, these results show that the likelihood estimation works well in estimating α when true values of α is small. For large values, the estimates appear to be positively biased. But, this biasness in the estimation of α does not create any practical problems in estimating other such as regression parameters. This is because, the proportion of missing data remains almost the same for any large values of α greater than 3, say. Thus, using $\hat{\alpha} = 3$ for $\alpha = 3$ or 4 or more will not make any difference in the estimates for other parameters.

Chapter 4

Complex Survey Based Incomplete Longitudinal Binary Models

In Chapter 2, longitudinal data subject to MAR was collected from K independent individuals. These individuals are considered to be the subjects or elements of a simple random sample (SRS) from an infinite population. In practice, specially in socio-economic research, it is most likely that the sample would be chosen from a finite population as opposed to an infinite population. Usually, stratified or two stage cluster sampling is adopted to select such a sample from the finite population. For example, Statistics Canada surveyed a labor and income dynamics (SLID) data set from 1993 to 1998 [Sutradhar and Kovacevic (2000)] on unemployment data recorded from a sample of size more than 35,000, where initially these individuals were selected based on a suitable complex such as cluster sampling. In this chapter we consider this finite sampling issue in the context of incomplete longitudinal data analysis. That is, in any inferences we examine the effects of

(1) design weights assigned for an individual to be selected in the sample,

(2) longitudinal correlation structure of the selected individuals, and

(3) missing mechanism that determines the availability of the responses over a period of time for the selected individuals.

Note that when missingness was exploited for the longitudinal binary responses, it was demonstrated in Chapter 2 that the response y_{it} follows a binary distribution with combined (missingness and longitudinal correlations) probability μ^*_{it}, i.e.,

$$y_{it} \sim bin(\mu^*_{it}), \;\; \mu^*_{it} = \lambda_{it|t-1}(\beta, \rho) w_{it}(\alpha). \tag{4.1}$$

We now consider an additional sampling design issue to be used to select the i-th individual from a finite population of size N whereas the i-th individual was selected based on a simple random sampling under the binary model discussed in Chapter 2 and count data model discussed in Chapter 3. For this purpose, we show below how to construct a stratified random sampling (StRS) scheme based design weight to be associated with an individual (say i-th individual).

4.1 Construction of the Survey Design Weights

Let there be N individuals in a finite population. Also let $\tilde{x}_i = (\tilde{x}_{i1}, \cdots, \tilde{x}_{iv}, \cdots, \tilde{x}_{il})'$ be a vector of l prognostic covariates. In the end, we are interested to find the effects of the diagnostic covariates $x_{it} = (x_{it1}, \cdots, x_{itp})'$ as introduced in Chapters 2 and 3. Note that by nature, $\tilde{x}_i$ and x_{it} are two different sets of covariates. Also note that to determine the design weights for the i-th individual we need to consider the prognostic covariates $\tilde{x}_i$ before the longitudinal study gets started. Thus $\tilde{x}_i$ is time independent.

Suppose that the prognostic covariates $\tilde{x}_i$ are known for all $i = 1, \cdots, N$, N being the finite population size, whereas we will still use K as the sample size which is a subset of N, K being much smaller as compared to N, i.e., $K << N$. Further suppose that $\tilde{x}_{iv}$ $(v = 1, \cdots, l)$ can be categorized into c_v categories. Thus, $\tilde{x}_{i1}$ has c_1 levels and similarly $\tilde{x}_{il}$ has c_l levels. Let L denote the number of strata based on these categories of the prognostic covariates. That is

$$L = \prod_{v=1}^{l} c_v. \tag{4.2}$$

Next, suppose that N individuals are distributed to these L strata such that N_h represents the size of the h-th $(h = 1, \cdots, L)$ stratum so that $\sum_{h=1}^{L} N_h = N$. Now as opposed to the infinite population, K individuals will be chosen from N individuals such that the identity of the selected individual will also play a role, such as the individual may belong to any of the L strata. Let $\sum_{h=1}^{L} n_h = K$, where n_h is the size of the sample of individuals that belong to the h-th stratum. In general n_h is determined by proportional allocation so that

$$\begin{aligned} n_h \propto N_h & \\ \Rightarrow n_h &= c_0 N_h, \text{ for a normalizing constant } C_0 \\ \Rightarrow \sum_{h=1}^{L} n_h &= c_0 \sum_{h=1}^{L} N_h \\ \Rightarrow K &= c_0 N \\ \Rightarrow c_0 &= \frac{K}{N} \\ \Rightarrow n_h &= (K/N) N_h. \end{aligned} \tag{4.3}$$

Now, if the i-th individual is known to belong to the h-th stratum (with size N_h), in the finite population, then this individual has the probability (n_h/N_h) for his/her inclusion in the h-th stratum of the sample. We denote this probability by $p_{i(h)}$, and write

$$p_{i(h)} = \frac{n_h}{N_h}. \tag{4.4}$$

It would, therefore, be efficient to incorporate this additional strata related cross-sectional information for the i-th individual to make any inferences for the longitudinal data to be collected subject to MAR. For this purpose, we now define the design weight for this i-th individual with inclusion probability $p_{i(h)} = n_h/N_h$. The design weight, $w^d_{i(h)}$ say, for the i-th individual in the h-th stratum is the inverse of its inclusion probability $p_{i(h)}$, i.e.,

$$w^d_{i(h)} = \frac{1}{p_{i(h)}} = \frac{N_h}{n_h} \tag{4.5}$$

so that

$$\sum_{h=1}^{L} \sum_{i=1}^{n_h} w^d_{i(h)} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{N_h}{n_h} = \sum_{h=1}^{L} N_h = N. \tag{4.6}$$

Thus, the design weight for an individual represents the ratio of the number of similar individuals in the finite population as compared to the sample. For example, if i-th individual in the h-th stratum has inclusion probability 1 in 20 (i.e. $p_{i(h)} = 1/20$), then this individual represents on the average 20 similar individuals of the finite population and his design weight would be 20, i.e. $w^{d}_{i(h)} = 20$.

4.2 Proposed Estimating Equations for the Survey Data

4.2.1 For survey based complete longitudinal data

Suppose that $y_{i(h)} = (y_{i1(h)}, \cdots, y_{iT(h)})'$ represents the $T \times 1$ complete repeated response vector from the i-th individual given that the individual belongs to the h-th ($h = 1, \cdots, L$) stratum in the finite/survey population. Also suppose that $\mu_{i(h)} = E(Y_{i(h)})$ and $\Sigma_{i(h)} = cov(Y_{i(h)})$ be the model based mean vector and covariance matrix of $y_{i(h)}$, respectively. If all responses $y_{i(h)}$ for $i = 1, \cdots, N_h$ and $h = 1, \cdots, L$, were known, then following (1.24), one would have estimated the regression effects β by solving

$$\sum_{h=1}^{L}\sum_{i=1}^{N_h} \frac{\partial \mu'_{i(h)}}{\partial \beta} \Sigma^{-1}_{i(h)} [y_{i(h)} - \mu_{i(h)}] = 0. \tag{4.7}$$

But, as $\sum_{h=1}^{L} N_h = N$ is very large (an unmanageable size from practical point of view), it is appropriate to choose a sample of $K << N$ individuals with $\sum_{h=1}^{L} n_h = K$, n_h being the size of the h-th stratum in the sample. If the repeated responses are now observed from these $K = \sum_{h=1}^{L} n_h$ individuals, the sample based estimating equation corresponding to (4.7) is given by

$$\sum_{h=1}^{L}\sum_{i=1}^{n_h} w^{d}_{i(h)} \frac{\partial \mu'_{i(h)}}{\partial \beta} \Sigma^{-1}_{i(h)} [y_{i(h)} - \mu_{i(h)}] = 0, \tag{4.8}$$

where $w^{d}_{i(h)}$ (4.5) represents the design weight for the i-th individual belonging to the h-th stratum. Let s^{**} denote the selected sample of $K = \sum_{h=1}^{L} n_h$ individuals. For convenience, the estimating equation in (4.8) may be re-expressed as

$$\sum_{i\in s^{**}} w^d_{is^{**}} \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1}(y_i - \mu_i) = 0 \tag{4.9}$$

[Binder (1983), Sutradhar and Kovacevic (2000)].

Note that all vectors and matrices such as $y_{i(h)}$, $\Sigma_{i(h)}$ and Σ_i used above are written for complete data set-up without any additional notation for the 'completeness'. To indicate the complete data situation, they could however be written as $y^c_{i(h)}$, $\Sigma^c_{i(h)}$ and Σ^c_i, respectively. But, we avoided this notation, as for the incomplete data in the next section, we denote these quantities by $\tilde{y}_{i(h)}$, $\tilde{\Sigma}_{i(h)}$ and $\tilde{\Sigma}_i$ respectively, indicating the dimension adjustment for missingness.

4.2.2 CWGQL estimation for survey based incomplete longitudinal data

Note that if the repeated data collected from all individuals in the finite population are subject to MAR, then following (2.34), one could have estimate β by solving the proposed CWGQL estimating equation

$$\sum_{h=1}^{L}\sum_{i=1}^{N_h} \frac{\partial \lambda_{i(h)}'}{\partial \beta} [\tilde{\Sigma}_{i(h)w}]^{-1}[\tilde{y}_{i(h)} - \lambda_{i(h)}] = 0. \tag{4.10}$$

Now, if a sample of size $K = \sum_{h=1}^{L} n_h$ is selected from the finite population of size N and the repeated responses from these selected individuals are subject to MAR, then the design based CWGQL (DBCWGQL) has the form

$$\sum_{h=1}^{L}\sum_{i=1}^{n_h} w^d_{i(h)} \frac{\partial \lambda_{i(h)}'}{\partial \beta} [\tilde{\Sigma}_{i(h)w}]^{-1}[\tilde{y}_{i(h)} - \lambda_{i(h)}] = 0, \tag{4.11}$$

which, following (4.9), may be re-expressed as

$$\sum_{i\in s^{**}} w^d_{is^{**}} \frac{\partial \lambda_i'}{\partial \beta} [\tilde{\Sigma}_{iw}]^{-1}[\tilde{y}_i - \lambda_i] = 0, \tag{4.12}$$

Note that the estimating equation (4.12) shows the nature of complex sampling used for the sample selection. To be specific, s^{**} indicates that a StRS [see also (4.9)] is used for the selction of $K = \sum_{h=1}^{L} n_h$ individuals. The equations (4.11) and (4.12) are exactly the same, but written in two different forms. Further note that if the sample is chosen by using other sampling design, it would be easy to use the form in (4.12) by replacing s^{**} with appropriate sample such as s^*, say, for the SRS case.

4.2.3 A sampling design based simulation study for incomplete longitudinal binary data

(a) Finite population structure

In the simulation study, we consider $N = 500$ individuals in the finite population. Suppose that the education of the individual and the family income of all these 500 individuals are known in advance before the longitudinal study. Further suppose that family income ($\tilde{x}_1$) has $c_1 = 3$ levels (High$\equiv H_I$, Medium $\equiv M_I$, Low $\equiv L_I$) and education ($\tilde{x}_2$) has $c_2 = 2$ levels (High $\equiv H_E$, Low $\equiv L_E$). It is clear that the i-th ($i = 1, \cdots, N$) individual belongs to one of the $L = c_1 \times c_2 = 6$ strata. Next, suppose that the marginal probabilities are known as $P(\tilde{x}_1 \in L_I) = 0.3$, $P(\tilde{x}_1 \in M_I) = 0.6$, $P(\tilde{x}_1 \in H_I) = 0.1$, $P(\tilde{x}_2 \in H_E) = 0.3$, and $P(\tilde{x}_2 \in L_E) = 0.7$. This leads to the probability table given in Table 4.1. Consequently, as a part of our design plan, the distribution of 500 individuals to 6 strata is known. For example, there are $500 \times 0.21 = 105$ individuals in the first stratum under the finite population. Suppose that we consider a sample of size $K = 100$ individuals. These 100 individuals will be selected from the stratified population (specified by Table 4.1) according to a desired sampling scheme, such as stratified random sampling (StRS). A simple random sampling (SRS) also will be considered and the inferences based on SRS and StRS will be compared.

(b) Simple random sampling

Here we select a sample of size $K = 100$ from $N = 500$ using SRS. Thus, we ignore the strata in the finite population and pretend that the population of 500 individuals

Table 4.1: An experimental stratified population

Stratum	1	2	3	4	5	6
Stratum ID ($\tilde{x}_1\tilde{x}_2$)	L_IL_E	L_IH_E	M_IL_E	M_IH_E	H_IL_E	H_IH_E
Probability	0.21	0.09	0.42	0.18	0.07	0.03

is homogeneous. To do this, we generate a sample of 100 values from the uniform distribution and select the individuals based on these uniform values (in integer) from 500 individuals. Thus, the sample is chosen by using SRS with replacement. Note that the selected individuals are likely to vary from simulation to simulation.

(c) Stratified random sampling

Here we select a SRS of size n_h with replacement from N_h individuals in the h-th stratum under the finite population. Thus, under StRS scheme, we select $\sum_{h=1}^{L} n_h = K = 100$ individuals from the stratified finite population of size $\sum_{h=1}^{L} N_h = N = 500$. Sample size n_h for the h-th $(h = 1, \cdots, L)$ stratum is determined by proportional allocation (4.3). Thus, when we select an individual in the sample by using the StRS, the identity of the individual's stratum is known.

(d) Incomplete longitudinal data generation for the selected sample

Note that after selecting the sample of size $K = 100$ either based on SRS or StRS, we then start the data collection longitudinally over a period of short time $T = 4$. This means, the desired repeated responses (binary or count) along with a set of diagnostic covariates will be collected during the period of the study. Here it is expected that a few longitudinal responses are likely to be missing following the MAR mechanism. As far as diagnostic covariates are concerned, we consider, for example, $p = 3$ covariates namely, age (x_{it1}), gender (x_{it2}) and smoking status (x_{it3}) of the i-th individual. The values of these 3 covariates are specified as follows.

(i) For x_{it1}, we generate x_{i11} $(t = 1)$ from a uniform distribution within the range from 18 and 65, for all $i = 1, \cdots, K = 100$. It then follows that the values of x_{it1} for $t = 2, \cdots, 4$ are known. Note that in the simulations, we use $(x_{it1} - 40)$ for x_{it1}, to avoid large numerical values for the original age x_{it1}.

(ii) Note that x_{it2} is however a fixed covariate. We generate this gender covariate from a binary distribution with selection probability 0.5, i.e. $x_{it2} \sim bin(0.5)$.

(iii) In practice, smoking status of the selected individual will be collected from the individual over the longitudinal duration of the study. In the simulation study, we assume that $x_{it3} \sim bin(0.3)$ for $t = 1, 2$. For the remaining time points i.e., for $t = 3$ and 4, we choose x_{it3} as $x_{it3} \sim bin(0.25)$.

For a selected individual, we now generate the incomplete repeated binary responses $y_{i1}, \cdots, y_{iT_i}$ following the conditional probability model (2.11) for longitudinally MAR data. The generation of this type of incomplete longitudinal binary data is given in detail in Section 2.6.1. Recall that when $R_{it} = 1$, the binary response y_{it} is generated following (2.11) [see also (4.1)], i.e., by using

$$y_{it} \sim bin[\lambda_{it|t-1}(\beta, \rho) w_{it}(\alpha)],$$

where $\lambda_{it|t-1}(\beta, \rho) = \mu_{it} + \rho(y_{i,t-1} - \mu_{i,t-1})$. Note that since the finite population consists of strata, the finite population model based marginal mean of y_{it} will also be affected by the stratification. Thus, we write

$$\mu_{it}(\gamma, \beta) = \frac{exp[(\gamma_1\delta_{i1} + \cdots + \gamma_5\delta_{i5}) + (x_{it1}\beta_1 + x_{it2}\beta_2 + x_{it3}\beta_3)]}{1 + exp[(\gamma_1\delta_{i1} + \cdots + \gamma_5\delta_{i5}) + (x_{it1}\beta_1 + x_{it2}\beta_2 + x_{it3}\beta_3)]}, \tag{4.13}$$

where

$$\delta_{ij} = 0, \text{ for all } j = 1, \cdots, 5, \tag{4.14}$$

is used to indicate that the i-th individual, whether selected by SRS or StRS, belongs to the 6-th $(H_I H_E)$ stratum (usually called as the reference group), and

$$\delta_{ij} = 1, \ \delta_{il} = 0 \ (l \neq j, \ j, l = 1, \cdots, 5) \tag{4.15}$$

indicates that the i-th individual, whether selected by SRS or StRS, belongs to the j-th ($j = 1, \cdots, 5$) stratum. In (4.13), γ and β are used to represent the effects of prognostic and diagnostic covariates. That is, $\gamma = (\gamma_1, \cdots, \gamma_{L-1})'$ with $L = 6$ and $\beta = (\beta_1, \cdots, \beta_p)'$ with $p = 3$. With regard to the true values of $\gamma = (\gamma_1, \cdots, \gamma_5)'$ and $\beta = (\beta_1, \beta_2, \beta_3)'$ in (4.13), we consider

$$\gamma_1 = 1,\ \gamma_2 = 1,\ \gamma_3 = 0.5,\ \gamma_4 = 0.5,\ \gamma_5 = 0.25, \text{ and}$$

$$\beta_1 = 0,\ \beta_2 = 0,\ \beta_3 = 0.$$

Note that as far as the value of $w_{it}(\alpha)$ is concerned, it is calculated in the same way as in (2.10) i.e.,

$$\begin{aligned} w_{it}(\alpha) &= \prod_{j=1}^{t} g_{ij}(\alpha \mid y_{i,j-1}) \\ &= \prod_{j=1}^{t} \frac{exp(1 + \alpha y_{i,j-1})}{1 + exp(1 + \alpha y_{i,j-1})}. \end{aligned} \tag{4.16}$$

Furthermore, the past responses involved in (4.16) are now generated following (4.1) [see also (2.11)] with μ_{it} as in (4.13).

Note that as far as the missing mechanism is concerned, it may happen in practice that while MAR is an appropriate mechanism for one stratum, MCAR, for example, could be appropriate for another stratum, and so on. But, for simplicity, we have assumed the same missing mechanism under all strata.

(e) Estimation of parameters

In each simulation, we select a random sample of size $K = 100$ from the finite population of size $N = 500$ either by a SRS or by a StRS. We then generate $\sum_{i=1}^{K} T_i$ longitudinally MAR binary responses from the selected individuals. Note that we have used $\alpha = 5$, α being the dependence parameter of the non-response model (2.7), to generate such incomplete data. To represent longitudinal correlations, we consider $\rho = 0.2$, 0.4, 0.6 and 0.8. Following the data generation, we then estimate the effects of diagnostic and/or prognostic covariates by the proposed DBCWGQL (design based conditional weighted generalized quasilikelihood) approach (4.12). The correlation index parameter ρ involved in model (4.1) is estimated by two moment equations, namely WMM (2.37) and MM (2.36).

(i) SRS

In any SRS based inferences, the existence of strata in the finite population is completely ignored. This is why, here, we pretend that the binary outcome is affected only by the set of diagnostic covariates and accordingly we only estimate β_1, β_2 and β_3 by the proposed DBCWGQL estimating equation (4.12). To be more specific, even though the incomplete longitudinal responses for the i-th individual in SRS were generated by using $\mu_{it}(\gamma, \beta)$ (4.13), we however estimate β only and thus the estimating equation now uses $\mu_{it}(\beta)$, i.e.,

$$\mu_{it} = \mu_{it}(\beta) = \frac{exp[(x_{it1}\beta_1 + x_{it2}\beta_2 + x_{it3}\beta_3)]}{1 + exp[(x_{it1}\beta_1 + x_{it2}\beta_2 + x_{it3}\beta_3)]}. \tag{4.17}$$

Specifically, the estimating equation (4.12) for β under the SRS may be expressed as

$$\sum_{i \in s^*} w^d_{is^*} \frac{\partial \lambda_i'(\beta)}{\partial \beta} [\tilde{\Sigma}_{iw}(\beta)]^{-1} [\tilde{y}_i - \lambda_i(\beta)] = 0, \tag{4.18}$$

where s^*, as indicated earlier, represents the K individuals chosen by SRS. Furthermore, in (4.18), $\lambda_i(\beta)$ and $\tilde{\Sigma}_{iw}(\beta)$ are written from (2.34) by replacing μ_{it} with $\mu_{it}(\beta)$ (4.17), and $w^d_{is^*} = 1/p_i$, $p_i = K/N$ being the SRS based inclusion probability for the i-th individual. Note that the SRS based DBCWGQL estimating equation (4.18) appears to have the same form as that of the CWGQL estimating equation (2.34) under the infinite population set-up. But, there is a big difference in the interpretation of the parameters involved in these equations. To be specific, $\tilde{y}_i$ in (2.34) is chosen by SRS from a true homogeneous population, whereas $\tilde{y}_i$ in (4.18) is chosen by SRS from a stratified finite population. Thus, λ_i and $\tilde{\Sigma}_{iw}$ in (2.34) are the true model parameters, whereas $\lambda_i(\beta)$ and $\tilde{\Sigma}_{iw}(\beta)$ in (4.18) are 'working' parameters. This is because when finite population contains several strata, λ_i should be the function of $\mu_{it}(\gamma, \beta)$ (4.13). This is further discussed below under (ii) StRS. It now follows that the mis-specified sampling design (SRS) based estimating equation (4.18) is not unbiased for zero. Thus, the solution of (4.18) for β is expected to be biased. Our simulation results in this section also verify this biasness.

Further note that the estimation of β by (4.18) requires the estimate of ρ. These two parameters are estimated iteratively. For the estimation of ρ parameter (as a function of β), we consider the MM (2.36) and the WMM (2.37) approaches, but use $\mu_{it}(\beta)$ (4.17) for μ_{it} as we did in writing the estimating equation (4.18) for β.

Note that for the purpose of constructing confidence intervals for the β parameter, one may need the estimated standard errors (ESEs) of the components of $\hat{\beta}$. But, as (4.18), because of the use of SRS, is a biased equation for β, one can not obtain the $cov(\beta)$ directly by using this equation. Instead, we may obtain the matrix of mean squared errors (instead of covariance matrix) for the SRS based estimator of β as given by

$$
\begin{aligned}
MSE(\hat{\beta}) \mid_{SRS} &= \left[\sum_{i \in s^*} w^d_{is^*} \frac{\partial \lambda'_i(\beta)}{\partial \beta} [\tilde{\Sigma}_{iw}(\beta)]^{-1} \frac{\partial \lambda_i(\beta)}{\partial \beta'}\right]^{-1} \times \\
& \quad MSE\left[\sum_{i \in s^*} w^d_{is^*} \frac{\partial \lambda'_i(\beta)}{\partial \beta} [\tilde{\Sigma}_{iw}(\beta)]^{-1} (\tilde{Y}_i - \lambda_i(\beta))\right] \times \\
& \quad \left[\sum_{i \in s^*} w^d_{is^*} \frac{\partial \lambda'_i(\beta)}{\partial \beta} [\tilde{\Sigma}_{iw}(\beta)]^{-1} \frac{\partial \lambda_i(\beta)}{\partial \beta'}\right]^{-1} \\
&= \left[\sum_{i \in s^*} w^d_{is^*} \frac{\partial \lambda'_i(\beta)}{\partial \beta} [\tilde{\Sigma}_{iw}(\beta)]^{-1} \frac{\partial \lambda_i(\beta)}{\partial \beta'}\right]^{-1} \times \\
& \quad \Bigg[\sum_{i \in s^*} (w^d_{is^*})^2 \frac{\partial \lambda'_i(\beta)}{\partial \beta} [\tilde{\Sigma}_{iw}(\beta)]^{-1} E(\tilde{Y}_i - \lambda_i(\beta))(\tilde{Y}_i - \lambda_i(\beta))' \\
& \quad [\tilde{\Sigma}_{iw}(\beta)]^{-1} \frac{\partial \lambda_i(\beta)}{\partial \beta'}\Bigg] \left[\sum_{i \in s^*} w^d_{is^*} \frac{\partial \lambda'_i(\beta)}{\partial \beta} [\tilde{\Sigma}_{iw}(\beta)]^{-1} \frac{\partial \lambda_i(\beta)}{\partial \beta'}\right]^{-1}.
\end{aligned}
\tag{4.19}
$$

Next, this $MSE(\hat{\beta})$ may be decomposed as

$$
\begin{aligned}
MSE(\hat{\beta})\mid_{SRS} \;=\; & \left[\sum_{i\in s^*} w^d_{is^*}\frac{\partial\lambda_i'(\beta)}{\partial\beta}[\tilde{\Sigma}_{iw}(\beta)]^{-1}\frac{\partial\lambda_i(\beta)}{\partial\beta'}\right]^{-1}\times \\
& \left[\sum_{i\in s^*}(w^d_{is^*})^2\frac{\partial\lambda_i'(\beta)}{\partial\beta}[\tilde{\Sigma}_{iw}(\beta)]^{-1}\tilde{\Sigma}_{iw}(\gamma,\beta)[\tilde{\Sigma}_{iw}(\beta)]^{-1}\frac{\partial\lambda_i(\beta)}{\partial\beta'}\right]\times \\
& \left[\sum_{i\in s^*} w^d_{is^*}\frac{\partial\lambda_i'(\beta)}{\partial\beta}[\tilde{\Sigma}_{iw}(\beta)]^{-1}\frac{\partial\lambda_i(\beta)}{\partial\beta'}\right]^{-1} \\
+ \; & \left[\sum_{i\in s^*} w^d_{is^*}\frac{\partial\lambda_i'(\beta)}{\partial\beta}[\tilde{\Sigma}_{iw}(\beta)]^{-1}\frac{\partial\lambda_i(\beta)}{\partial\beta'}\right]^{-1}\left[\sum_{i\in s^*}(w^d_{is^*})^2\frac{\partial\lambda_i'(\beta)}{\partial\beta}[\tilde{\Sigma}_{iw}(\beta)]^{-1}\right. \\
& \left.\left\{[\lambda_i(\beta)-\lambda_i(\gamma,\beta)][\lambda_i(\beta)-\lambda_i(\gamma,\beta)]'\right\}[\tilde{\Sigma}_{iw}(\beta)]^{-1}\frac{\partial\lambda_i(\beta)}{\partial\beta'}\right]\times \\
& \left[\sum_{i\in s^*} w^d_{is^*}\frac{\partial\lambda_i'(\beta)}{\partial\beta}[\tilde{\Sigma}_{iw}(\beta)]^{-1}\frac{\partial\lambda_i(\beta)}{\partial\beta'}\right]^{-1}, \qquad (4.20)
\end{aligned}
$$

where $\tilde{\Sigma}_{iw}(\gamma,\beta) = E[\tilde{Y}_i - \lambda_i(\gamma,\beta)][\tilde{Y}_i - \lambda_i(\gamma,\beta)]'$ is the finite (stratified) population based covariance matrix. One may then re-write the $MSE(\hat{\beta})$ in (4.20) as

$$
MSE(\hat{\beta})\mid_{SRS} = cov(\hat{\beta})\mid_{SRS} + \left[bias(\hat{\beta})\right]\left[bias(\hat{\beta})\right]'\mid_{SRS}. \qquad (4.21)
$$

Note that one may now estimate the $cov(\hat{\beta})\mid_{SRS}$ by estimating the first term in (4.21). This may be achieved by replacing finite population based $\tilde{\Sigma}_{iw}(\gamma,\beta)$ in the first term of (4.20) with sampling design based $\tilde{\Sigma}_{iw}(\beta)$. Thus, the estimator for the covariance matrix of SRS based regression estimator is given by

$$
\hat{cov}(\hat{\beta})\mid_{SRS} = \left[\sum_{i\in s^*}\frac{\partial\lambda_i'(\beta)}{\partial\beta}[\tilde{\Sigma}_{iw}(\beta)]^{-1}\frac{\partial\lambda_i(\beta)}{\partial\beta'}\right]^{-1}. \qquad (4.22)
$$

Computational formulas under SRS

Note that we have considered a SRS of size K. So, for these K individuals belonging to the sample 's^*', we may write $w^d_{is^*} = 1/\pi_i = N/K$ as mentioned earlier, and the estimating equation (4.18), for convenience, may be re-written as

$$\sum_{i=1}^{K} \frac{\partial \lambda_i'(\beta)}{\partial \beta} [\tilde{\Sigma}_{iw}(\beta)]^{-1} [\tilde{y}_i - \lambda_i(\beta)]_{(s^*)} = 0. \tag{4.23}$$

Here $i = 1, \cdots, K$ refers to those individuals $i \in s^*$. By the same token, the estimated covariance matrix of $\hat{\beta}$ given by (4.22), may be computed by

$$\hat{cov}(\hat{\beta}) \mid_{SRS} = \left[\sum_{i=1}^{K} \frac{\partial \lambda_i'(\beta)}{\partial \beta} [\tilde{\Sigma}_{iw}(\beta)]^{-1} \frac{\partial \lambda_i(\beta)}{\partial \beta'} \right]^{-1}_{(s^*)} . \tag{4.24}$$

The simulated DBCWGQL estimates of β and WMM (2.37) estimate of ρ under the SRS scheme are given in columns 4 to 7 of Table 4.2. Table 4.3 exhibits similar results (as in Table 4.2) when ρ is estimated by MM (2.36). Since the confidence intervals for β may also be of interest, we produce the SSEs of the components of β in column 5. The ESEs computed from the diagonal elements of (4.24) are given in column 6 of Table 4.2 and 4.3. The ESEs appear to be reasonably close to the SSEs indicating that the performance of (4.24) is satisfactory.

Note that our main objective here is to examine the effects of the diagnostic covariates namely, age, gender and smoking status when the finite population consists of strata but the sample contains SRS based individuals. It is clear from both Tables 4.2 and 4.3 that as expected, the SRS based DBCWGQL or CWGQL estimation approach in general produces highly biased estimates. For example, Table 4.2 shows that when $\rho = 0.6$, SRS based estimates for the components of β are

$$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)' = (0.006, 0.646, 0.180)',$$

whereas the true values are

$$\beta = (\beta_1, \beta_2, \beta_3)' = (0, 0, 0)'.$$

Here $\hat{\beta}_2$ and $\hat{\beta}_3$ are highly biased. Similar conclusion may be drawn based on the results in Table 4.3. This biasness or inconsistency is not surprising because the finite population consists of strata but the SRS scheme completely ignores these strata information while making inferences for the parameters. Note that this biasness, in fact, may be understood from the second term in (4.20). Thus SRS based regression estimates can not be trusted when the population contains individuals under various strata specially when the variability between strata is large.

As far as the performance of ρ estimator is concerned, the SRS based WMM or MM estimates of ρ, in general, appear to be satisfactory, even though estimates of β were found to be highly biased. When the estimates are closely looked at, the WMM approach appears to underestimate and the MM approach appears to overestimate the ρ parameter. But, the amount of biases are not so large.

(ii) StRS

Unlike the SRS, under the stratified sampling scheme, the stratum identity of the selected individuals are kept in tact while estimating the parameters. Thus, the estimating equation for β has the form

$$\sum_{i\in s^{**}} w^d_{is^{**}} \frac{\partial\lambda_i^{'}(\gamma,\beta)}{\partial\beta}[\tilde{\Sigma}_{iw}(\gamma,\beta)]^{-1}[\tilde{y}_i - \lambda_i(\gamma,\beta)] = 0, \tag{4.25}$$

which is different than (4.18). This is because $\lambda_i(\gamma,\beta)$ and $\tilde{\Sigma}_{iw}(\gamma,\beta)$ in (4.25) are functions of $\mu_{it}(\gamma,\beta)$ (4.13), whereas $\lambda_i(\beta)$ and $\tilde{\Sigma}_{iw}(\beta)$ in (4.18) are functions of $\mu_{it}(\beta)$ (4.17). In fact, it is now also possible to estimate γ_r $(r = 1,\cdots,L-1)$ parameters by using an estimating equation similar to (4.25). More specifically, let $\theta = (\gamma^{'},\beta^{'})^{'}$, where γ's are effects of strata (prognostic covariates) and β's are the effects of diagnostic covariates, and this θ parameter may be estimated by using

$$\sum_{i\in s^{**}} w^d_{is^{**}} \frac{\partial\lambda_i^{'}(\theta)}{\partial\theta}[\tilde{\Sigma}_{iw}(\theta)]^{-1}[\tilde{y}_i - \lambda_i(\theta)] = 0. \tag{4.26}$$

Note that our main objective is to compare the estimates of $\beta = (\beta_1,\cdots,\beta_p)^{'}$ obtained by the SRS and the StRS. We however use (4.26) to obtain joint estimates for the components of θ under the StRS and compare β estimates with the corresponding β estimates obtained by solving the SRS based estimating equation (4.18).

Further note that since (4.26) is an unbiased estimating equation for zero, the covariance of $\hat{\theta}$ has the formula

$$\begin{aligned} cov(\hat{\theta}) &= \left[\sum_{i\in s^{**}} w^d_{is^{**}} \frac{\partial \lambda_i'(\theta)}{\partial \theta}[\tilde{\Sigma}_{iw}(\theta)]^{-1}\frac{\partial \lambda_i(\theta)}{\partial \theta'}\right]^{-1} \times \\ & \quad cov\left[\sum_{i\in s^{**}} w^d_{is^{**}} \frac{\partial \lambda_i'(\theta)}{\partial \theta}[\tilde{\Sigma}_{iw}(\theta)]^{-1}[\tilde{Y}_i - \lambda_i(\theta)]\right] \times \\ & \quad \left[\sum_{i\in s^{**}} w^d_{is^{**}} \frac{\partial \lambda_i'(\theta)}{\partial \theta}[\tilde{\Sigma}_{iw}(\theta)]^{-1}\frac{\partial \lambda_i(\theta)}{\partial \theta'}\right]^{-1} \\ &= \left[\sum_{i\in s^{**}} w^d_{is^{**}} \frac{\partial \lambda_i'(\theta)}{\partial \theta}[\tilde{\Sigma}_{iw}(\theta)]^{-1}\frac{\partial \lambda_i(\theta)}{\partial \theta'}\right]^{-1} cov\left[\sum_{i\in s^{**}} w^d_{is^{**}}\tilde{z}_{is^{**}}\right] \times \\ & \quad \left[\sum_{i\in s^{**}} w^d_{is^{**}} \frac{\partial \lambda_i'(\theta)}{\partial \theta}[\tilde{\Sigma}_{iw}(\theta)]^{-1}\frac{\partial \lambda_i(\theta)}{\partial \theta'}\right]^{-1}, \end{aligned} \tag{4.27}$$

where $\tilde{z}_{is^{**}} = (\partial \lambda_i'(\theta)/\partial\theta)[\tilde{\Sigma}_{iw}(\theta)]^{-1}[\tilde{y}_i - \lambda_i(\theta)]$. Note that the estimation of this $cov(\hat{\theta})$ in (4.27) requires the estimation of $cov\left[\sum_{i\in s^{**}} w^d_{is^{**}}\tilde{z}_{is^{**}}\right]$. It would be however convenient to simplify this formula by re-expressing $\sum_{i\in s^{**}} w^d_{is^{**}}\tilde{z}_{is^{**}}$ in such way that one recognizes the stratum identity for $\tilde{z}_{is^{**}}$ under the finite population.

Computational formulas under StRS

For the computational purpose, it is now important to identify the stratum for the i-th individual selected in the sample 's^{**}' shown by (4.26). Since the individuals in s^{**} are chosen by StRS, $w^d_{is^{**}}$ in (4.26) may be expressed as

$$w^d_{is^{**}} \equiv w^d_{i(h)} = \frac{N_h}{n_h}, \text{ for } i = 1, \cdots, n_h$$

so that $\sum_{h=1}^{L} n_h = K$. Similarly λ_i and $\tilde{\Sigma}_{iw}$ for $i \in s^{**}$ in (4.26) may be expressed by $\lambda_{i(h)}(\theta)$ and $\tilde{\Sigma}_{i(h)w}(\theta)$, respectively. Thus, the estimating equation (4.26) has the simplified computational formula given by

$$\sum_{h=1}^{L}\frac{N_h}{n_h}\sum_{i=1}^{n_h}\frac{\partial \lambda_{i(h)}'(\theta)}{\partial\theta}[\tilde{\Sigma}_{i(h)w}(\theta)]^{-1}[\tilde{y}_{i(h)} - \lambda_{i(h)}(\theta)] = 0 \tag{4.28}$$

which is constructed by referring the strata under the finite population.

We now write a computational formula for the $\hat{cov}(\hat{\theta})$ in (4.27). In the fashion similar to (4.28), the first/third term in (4.27) may be written as

$$\left[\sum_{i\in s^{**}} w^d_{is^{**}} \frac{\partial\lambda_i'(\theta)}{\partial\theta}[\tilde{\Sigma}_{iw}(\theta)]^{-1}\frac{\partial\lambda_i(\theta)}{\partial\theta'}\right]^{-1} = \left[\sum_{h=1}^{L} \frac{N_h}{n_h}\sum_{i=1}^{n_h} \frac{\partial\lambda'_{i(h)}(\theta)}{\partial\theta}[\tilde{\Sigma}_{i(h)w}(\theta)]^{-1}\frac{\partial\lambda_{i(h)}(\theta)}{\partial\theta'}\right]^{-1}. \tag{4.29}$$

Next, we write a computational formula for the middle term $cov\left[\sum_{i\in s^{**}} w^d_{is^{**}}\tilde{z}_{is^{**}}\right]$ in (4.27) as follows:

Re-express $\tilde{z}_{is^{**}}$ in (4.27) as

$$\tilde{z}_{is^{**}} = \tilde{z}_{i(h)} = \frac{\partial\lambda'_{i(h)}(\theta)}{\partial\theta}[\tilde{\Sigma}_{i(h)w}(\theta)]^{-1}\left[\tilde{y}_{i(h)} - \lambda_{i(h)}(\theta)\right],$$

provided that i-th individual in the sample belongs to the h-th stratum. Let

$$z_{i(h)} = w^d_{i(h)}\tilde{z}_{i(h)}, \tag{4.30}$$

where $w^d_{i(h)} = N_h/n_h$. It then follows that

$$\begin{aligned} cov\left(\sum_{i\in s^{**}} w^d_{is^{**}}\tilde{z}_{is^{**}}\right) &= cov\left(\sum_{h=1}^{L}\sum_{i=1}^{n_h} z_{i(h)}\right) \\ &= cov\left(\sum_{h=1}^{L} n_h\bar{z}_{(h)}\right), \end{aligned} \tag{4.31}$$

where $\bar{z}_{(h)} = \sum_{i=1}^{n_h} z_{i(h)}/n_h : \ (p+L-1)\times 1$. Since L strata are independent, the covariance in (4.31) is written as

$$\begin{aligned} cov\left(\sum_{i\in s^{**}} w^d_{is^{**}}\tilde{z}_{is^{**}}\right) &= \sum_{h=1}^{L} n_h^2 cov(\bar{z}_{(h)}) \\ &= \sum_{h=1}^{L} n_h^2\frac{(N_h-n_h)}{N_h}\frac{S_{(h)}}{n_h}, \end{aligned} \tag{4.32}$$

where $S_{(h)} = \sum_{i=1}^{N_h}\left(z_{i(h)} - \bar{Z}_{(h)}\right)\left(z_{i(h)} - \bar{Z}_{(h)}\right)'/(N_h-1)$ with $\bar{Z}_{(h)} = \sum_{i=1}^{N_h} z_{i(h)}/N_h$. Since $S_{(h)}$ can be unbiasedly estimated with $s_{(h)}$ given by

$$\hat{S}_{(h)} = s_{(h)} = \frac{\sum_{i=1}^{n_h}\left(z_{i(h)} - \bar{z}_{(h)}\right)\left(z_{i(h)} - \bar{z}_{(h)}\right)'}{n_h - 1}, \tag{4.33}$$

we may then estimate the $cov\left(\sum_{i\in s^{**}} w^d_{is^{**}}\tilde{z}_{is^{**}}\right)$ in (4.32) as

$$
\begin{aligned}
\hat{cov}\left(\sum_{i\in s^{**}} w^d_{is^{**}}\tilde{z}_{is^{**}}\right) &= \sum_{h=1}^{L}\left(\frac{N_h - n_h}{N_h}\right) n_h s_{(h)} \\
&\approx \sum_{h=1}^{L}\frac{n_h}{n_h - 1}\sum_{i=1}^{n_h}(z_{i(h)} - \bar{z}_{(h)})(z_{i(h)} - \bar{z}_{(h)})', \qquad (4.34)
\end{aligned}
$$

when the sampling fraction n_h/N_h is negligible. Therefore, using (4.29) and (4.34) in (4.27), the covariance matrix of $\hat{\theta}$ may be estimated under the StRS as

$$
\begin{aligned}
\hat{cov}(\hat{\theta}) = &\left[\sum_{h=1}^{L}\frac{N_h}{n_h}\sum_{i=1}^{n_h}\frac{\partial\lambda'_{i(h)}(\theta)}{\partial\theta}[\tilde{\Sigma}_{i(h)w}(\theta)]^{-1}\frac{\partial\lambda_{i(h)}(\theta)}{\partial\theta'}\right]^{-1} \times \\
&\left[\sum_{h=1}^{L}\frac{n_h}{n_h - 1}\sum_{i=1}^{n_h}(z_{i(h)} - \bar{z}_h)(z_{i(h)} - \bar{z}_h)'\right] \times \\
&\left[\sum_{h=1}^{L}\frac{N_h}{n_h}\sum_{i=1}^{n_h}\frac{\partial\lambda'_{i(h)}(\theta)}{\partial\theta}[\tilde{\Sigma}_{i(h)w}(\theta)]^{-1}\frac{\partial\lambda_{i(h)}(\theta)}{\partial\theta'}\right]^{-1}. \qquad (4.35)
\end{aligned}
$$

After estimating the θ parameter, we use them to estimate the correlation index parameter ρ by a suitable moment approach. Note that when we estimate ρ under StRS, we have computed them for each stratum and then took their average. For example, suppose that $\hat{\rho}_{h,MM}$ is the estimate of ρ computed based on n_h selected individuals from the h-th stratum under the MM approach by using the formula (2.36). Then the MM estimate of ρ under StRS scheme becomes

$$
\hat{\rho}_{MM} = \frac{\sum_{h=1}^{L}\hat{\rho}_{h,MM}}{L}. \qquad (4.36)
$$

The simulation results for the DBCWGQL estimator of β obtained from (4.28) are given in the last 4 columns of Table 4.2 and Table 4.3, when ρ is estimated by the WMM (2.37) and the MM (2.36) approaches, respectively. It is clear from both tables that the DBCWGQL approach produces almost unbiased regression estimates. This is expected as the sample is selected from the finite stratified population using the appropriate StRS scheme, whereas SRS does not reflect the finite population nature.

For example, when ρ (=0.6) is estimated by MM, the StRS based regression estimates found in Table 4.3 are:

$$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)' = (0.001, 0.013, -0.001)'$$

for true values of

$$\beta = (\beta_1, \beta_2, \beta_3)' = (0, 0, 0)',$$

whereas under the SRS based β estimates were found to be

$$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)' = (0.004, 0.567, 0.102)'$$

showing large biases. Irrespective of the use of WMM or MM estimation for ρ, similar conclusions hold for any other DBCWGQL regression estimates. With regard to the estimation performance of ρ by WMM or MM approaches, the MM approach (2.36) was found to be better than the WMM approach (2.37) when both of them use strata based equation (4.36). For example, for $\rho = 0.6$, the WMM produces $\hat{\rho} = 0.464$ and the MM produces $\hat{\rho} = 0.567$.

When the regression estimates are compared under SRS and StRS schemes, StRS produces unbiased estimates for all the regression parameters, whereas the SRS produces highly biased estimates. As far as the standard errors of the regression estimates are concerned, the SRS produces biased regression estimates with smaller SSEs, as compared to the StRS based unbiased regression estimates with larger SSEs. This is in fact a serious problem for the SRS approach. This is because, small SSEs for biased regression estimates indicate that the estimates are almost always converging to the wrong values as compared to the true values of the parameters.

With regard to the estimation of SSE (i.e. true standard deviation of the estimator), it is found that the SRS based ESEs perform better than the StRS based ESEs in estimating the corresponding SSEs. One may, therefore, try to use alternative ESEs under the StRS scheme, but this is beyond the scope of the present study.

Table 4.2: Simulated sampling design based CWGQL (DBCWGQL) estimates for the diagnostic and/or prognostic covariates in MAR based incomplete longitudinal binary models with non-response index parameter $\alpha = 5$, and selected values of the longitudinal correlation ρ (estimated by WMM), based on 1000 simulations.

			Sampling Scheme							
			SRS				StRS			
ρ	Covariate	Parameter	SM	SSE	ESE	SMSE	SM	SSE	ESE	SMSE
0.2	Age	$\hat{\beta}_1$	0.005	0.012	0.010	0.000	0.000	0.011	0.010	0.000
	Male Vs Female	$\hat{\beta}_2$	0.548	0.217	0.215	0.348	0.015	0.299	0.274	0.090
	Smoker Vs Non	$\hat{\beta}_3$	0.305	0.272	0.251	0.167	-0.006	0.297	0.270	0.088
	L_IL_E Vs H_IH_E	$\hat{\gamma}_1$					1.088	0.364	0.350	0.141
	L_IH_E Vs H_IH_E	$\hat{\gamma}_2$					1.068	0.481	0.440	0.236
	M_IL_E Vs H_IH_E	$\hat{\gamma}_3$					0.553	0.297	0.263	0.091
	M_IH_E Vs H_IH_E	$\hat{\gamma}_4$					0.543	0.372	0.368	0.140
	H_IL_E Vs H_IH_E	$\hat{\gamma}_5$					0.363	0.732	0.615	0.548
		$\hat{\rho}_{WMM}$	0.186	0.078		0.006	0.140	0.110		0.016
0.4	Age	$\hat{\beta}_1$	0.005	0.012	0.011	0.000	0.000	0.012	0.011	0.000
	Male Vs Female	$\hat{\beta}_2$	0.597	0.244	0.231	0.416	-0.008	0.337	0.304	0.113
	Smoker Vs Non	$\hat{\beta}_3$	0.244	0.251	0.233	0.123	0.003	0.276	0.246	0.076
	L_IL_E Vs H_IH_E	$\hat{\gamma}_1$					1.123	0.416	0.383	0.188
	L_IH_E Vs H_IH_E	$\hat{\gamma}_2$					1.179	0.561	0.490	0.347
	M_IL_E Vs H_IH_E	$\hat{\gamma}_3$					0.583	0.312	0.285	0.104
	M_IH_E Vs H_IH_E	$\hat{\gamma}_4$					0.590	0.416	0.406	0.181
	H_IL_E Vs H_IH_E	$\hat{\gamma}_5$					0.363	0.795	0.679	0.845
		$\hat{\rho}_{WMM}$	0.345	0.082		0.010	0.295	0.120		0.025

(Table 4.2 contd...)

			Sampling Scheme							
			SRS				StRS			
ρ	Covariate	Parameter	SM	SSE	ESE	SMSE	SM	SSE	ESE	SMSE
0.6	Age	$\hat{\beta}_1$	0.006	0.014	0.012	0.000	0.002	0.014	0.013	0.000
	Male Vs Female	$\hat{\beta}_2$	0.646	0.283	0.249	0.498	0.018	0.379	0.337	0.144
	Smoker Vs Non	$\hat{\beta}_3$	0.180	0.223	0.205	0.082	-0.006	0.244	0.209	0.059
	L_IL_E Vs H_IH_E	$\hat{\gamma}_1$					1.193	0.481	0.426	0.269
	L_IH_E Vs H_IH_E	$\hat{\gamma}_2$					1.256	0.631	0.550	0.464
	M_IL_E Vs H_IH_E	$\hat{\gamma}_3$					0.624	0.361	0.314	0.146
	M_IH_E Vs H_IH_E	$\hat{\gamma}_4$					0.656	0.505	0.448	0.279
	H_IL_E Vs H_IH_E	$\hat{\gamma}_5$					0.454	0.940	0.778	0.925
		$\hat{\rho}_{WMM}$	0.515	0.084		0.014	0.464	0.117		0.032
0.8	Age	$\hat{\beta}_1$	0.006	0.015	0.013	0.000	0.003	0.017	0.014	0.000
	Male Vs Female	$\hat{\beta}_2$	0.683	0.317	0.275	0.567	0.006	0.440	0.378	0.194
	Smoker Vs Non	$\hat{\beta}_3$	0.078	0.176	0.156	0.037	-0.004	0.202	0.150	0.041
	L_IL_E Vs H_IH_E	$\hat{\gamma}_1$					1.261	0.539	0.486	0.359
	L_IH_E Vs H_IH_E	$\hat{\gamma}_2$					1.304	0.731	0.629	0.627
	M_IL_E Vs H_IH_E	$\hat{\gamma}_3$					0.683	0.408	0.348	0.200
	M_IH_E Vs H_IH_E	$\hat{\gamma}_4$					0.716	0.592	0.503	0.397
	H_IL_E Vs H_IH_E	$\hat{\gamma}_5$					0.537	1.029	0.880	1.142
		$\hat{\rho}_{WMM}$	0.715	0.074		0.013	0.653	0.106		0.033

Table 4.3: Simulated sampling design based CWGQL (DBCWGQL) estimates for the diagnostic and/or prognostic covariates in MAR based incomplete longitudinal binary models with non-response index parameter $\alpha = 5$, and selected values of the longitudinal correlation ρ (estimated by MM), based on 1000 simulations.

			Sampling Scheme							
			SRS				StRS			
ρ	Covariate	Parameter	SM	SSE	ESE	SMSE	SM	SSE	ESE	SMSE
0.2	Age	$\hat{\beta}_1$	0.004	0.011	0.011	0.000	0.000	0.011	0.010	0.000
	Male Vs Female	$\hat{\beta}_2$	0.490	0.209	0.227	0.284	0.013	0.293	0.276	0.086
	Smoker Vs Non	$\hat{\beta}_3$	0.210	0.249	0.236	0.106	-0.004	0.276	0.261	0.076
	$L_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_1$					1.004	0.357	0.352	0.127
	$L_I H_E$ Vs $H_I H_E$	$\hat{\gamma}_2$					0.990	0.475	0.442	0.226
	$M_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_3$					0.482	0.290	0.264	0.084
	$M_I H_E$ Vs $H_I H_E$	$\hat{\gamma}_4$					0.472	0.363	0.370	0.133
	$H_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_5$					0.301	0.710	0.616	0.506
		$\hat{\rho}_{MM}$	0.327	0.070		0.021	0.269	0.101		0.015
0.4	Age	$\hat{\beta}_1$	0.003	0.011	0.012	0.000	-0.001	0.012	0.011	0.000
	Male Vs Female	$\hat{\beta}_2$	0.525	0.234	0.243	0.331	-0.010	0.329	0.306	0.108
	Smoker Vs Non	$\hat{\beta}_3$	0.147	0.218	0.212	0.069	0.005	0.247	0.233	0.061
	$L_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_1$					1.024	0.406	0.385	0.165
	$L_I H_E$ Vs $H_I H_E$	$\hat{\gamma}_2$					1.081	0.549	0.492	0.308
	$M_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_3$					0.498	0.302	0.287	0.091
	$M_I H_E$ Vs $H_I H_E$	$\hat{\gamma}_4$					0.506	0.405	0.409	0.164
	$H_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_5$					0.296	0.773	0.680	0.600
		$\hat{\rho}_{MM}$	0.483	0.063		0.011	0.421	0.101		0.011

(Table 4.3 contd...)

			Sampling Scheme							
			SRS				StRS			
ρ	Covariate	Parameter	SM	SSE	ESE	SMSE	SM	SSE	ESE	SMSE
0.6	Age	$\hat{\beta}_1$	0.004	0.013	0.012	0.000	0.001	0.014	0.013	0.000
	Male Vs Female	$\hat{\beta}_2$	0.567	0.267	0.260	0.393	0.013	0.369	0.339	0.136
	Smoker Vs Non	$\hat{\beta}_3$	0.102	0.182	0.180	0.044	-0.001	0.209	0.196	0.044
	L_IL_E Vs H_IH_E	$\hat{\gamma}_1$					1.092	0.469	0.428	0.229
	L_IH_E Vs H_IH_E	$\hat{\gamma}_2$					1.158	0.616	0.552	0.405
	M_IL_E Vs H_IH_E	$\hat{\gamma}_3$					0.540	0.349	0.316	0.123
	M_IH_E Vs H_IH_E	$\hat{\gamma}_4$					0.574	0.488	0.451	0.244
	H_IL_E Vs H_IH_E	$\hat{\gamma}_5$					0.386	0.907	0.778	0.842
		$\hat{\rho}_{MM}$	0.634	0.060		0.005	0.567	0.097		0.010
0.8	Age	$\hat{\beta}_1$	0.004	0.014	0.013	0.000	0.002	0.016	0.014	0.000
	Male Vs Female	$\hat{\beta}_2$	0.605	0.298	0.282	0.455	0.006	0.431	0.379	0.186
	Smoker Vs Non	$\hat{\beta}_3$	0.035	0.129	0.134	0.018	-0.001	0.167	0.142	0.028
	L_IL_E Vs H_IH_E	$\hat{\gamma}_1$					1.177	0.533	0.488	0.316
	L_IH_E Vs H_IH_E	$\hat{\gamma}_2$					1.227	0.722	0.629	0.574
	M_IL_E Vs H_IH_E	$\hat{\gamma}_3$					0.613	0.393	0.349	0.167
	M_IH_E Vs H_IH_E	$\hat{\gamma}_4$					0.648	0.582	0.505	0.362
	H_IL_E Vs H_IH_E	$\hat{\gamma}_5$					0.489	1.007	0.878	1.070
		$\hat{\rho}_{MM}$	0.793	0.045		0.002	0.718	0.086		0.014

Chapter 5

Complex Survey Based Incomplete Longitudinal Models for Count Data

In this chapter we concentrate on the survey based incomplete longitudinal count data models as opposed to the survey based binary models discussed in the last chapter. Note however that there will be no differences in the forms for the estimating equations under the count data models as compared to the binary models. Thus, under SRS and StRS we may use the estimating equations (4.18) and (4.26) in estimating β respectively, by making appropriate changes in the formulas for $\mu_{it}(\gamma, \beta)$ and $\mu_{it}(\beta)$ as well as $\tilde{\Sigma}_{iw}(\gamma, \beta)$ and $\tilde{\Sigma}_{iw}(\beta)$. More specifically, $\mu_{it}(\beta)$ involved in $\lambda_i(\beta)$ in (4.18) is now given by

$$\mu_{it}(\beta) = exp\left(x_{it1}\beta_1 + x_{it2}\beta_2 + x_{it3}\beta_3\right), \tag{5.1}$$

and similarly $\mu_{it}(\gamma, \beta)$ involved in $\lambda_i(\theta)$ in (4.26) is given by

$$\mu_{it}(\gamma, \beta) = exp\left[(\gamma_1\delta_{i1} + \cdots + \gamma_5\delta_{i5}) + (x_{it1}\beta_1 + x_{it2}\beta_2 + x_{it3}\beta_3)\right]. \tag{5.2}$$

In view of the similarity between binary and Poisson cases, we do not re-produce the theoretical formulas for the estimation of β and γ under the count data models, whether SRS or StRS is chosen as the sampling design. Instead, in this chapter, we simply report the simulated estimates in Table 5.1 and 5.2 under the count data models, whereas similar estimates for the incomplete binary data were reported in Tables 4.2 and 4.3, respectively.

Note that in the simulation study, we have chosen the same design parameters as in the binary case. Thus, in the present simulation study, a sample (whether SRS or StRS based) of size $K = 100$ is chosen from a stratified finite population of size $N = 500$ containing $L = 6$ strata. Each of the selected individual provided T_i ($T_i \leq T,\ T = 4,\ i = 1, \cdots, K$) repeated count responses, the non-missing responses being determined based on MAR mechanism with non-response index parameter $\alpha = 5$ (approximately 94% response). We however remark that while the MAR mechanism for non-response indication and the sampling design for sample selection were almost the same for the binary and count data models, the longitudinal models for the repeated binary and count data are however different. To be specific, we have used the conditional linear binary dynamic (CLBD) model (1.26), namely

$$P(Y_{i1} = 1) = \mu_{i1},$$

$$P(Y_{it} = 1 \mid y_{i,t-1}) = \lambda_{i,t|t-1}(y_{i,t-1}) = \mu_{it} + \rho(y_{i,t-1} - \mu_{i,t-1}), \text{ for } t = 2, \cdots, T.$$

with $\mu_{it} = exp(x_{it}^{'}\beta)/[1 + exp(x_{it}^{'}\beta)]$ in conjunction with the MAR mechanism for the generation of the incomplete repeated binary data, whereas we have used the binomial thinning based AR(1) type dynamic relationship (1.21), namely,

$$\begin{aligned} y_{it} &= \rho * y_{i,t-1} + d_{it} \\ &= \sum_{j=1}^{y_{i,t-1}} b_j(\rho) + d_{it} \end{aligned}$$

in conjunction with the MAR mechanism to generate incomplete repeated count data.

The results in Table 5.1 and 5.2 show that the StRS based CWGQL approach performs very well in estimating both diagnostic and prognostic covariates effects. For example, for $\alpha = 5$, $\rho = 0.6$, the results in columns 11 and 7 in Table 5.2 show that the StRS based CWGQL approach estimates the diagnostic regression effects β_1, β_2

and β_3 with SMSEs 0.000, 0.016, and 0.006 respectively, whereas SRS based CWGQL approach produces the estimates with SMSEs as 0.000, 0.417 and 0.011. When the results in Table 5.2 are compared with those in 4.3 for the binary data, the estimation performance appears to be better under the count data models. This is because, the SMSEs are in general found to be smaller in Table 5.2 as compared to that of 4.3 in estimating both β and γ parameters. Similar results hold for Table 5.1 and 4.2. The longitudinal correlation estimates appear to be similar both for the binary and count data cases, the binary data based estimates being slightly better. Note however that our main purpose is not to compare the estimation performances under the binary and count data models. It is rather important to compare the performance of the proposed CWGQL approach as compared to the existing UWGEE and other similar approaches, which we have done under both binary and count models in Chapters 2 and 3, respectively in non-surveyed incomplete longitudinal set-up. As mentioned earlier, this comparison was not continued in Chapters 4 and 5, because of the reasons that the CWGQL was found to be uniformly better in Chapters 2 and 3.

Table 5.1: Simulated sampling design based CWGQL (DBCWGQL) estimates for the diagnostic and/or prognostic covariates in MAR based incomplete longitudinal models for count data with non-response index parameter $\alpha = 5$, and selected values of the longitudinal correlation ρ (estimated by WMM), based on 1000 simulations.

			Sampling Scheme							
			SRS				StRS			
ρ	Covariate	Parameter	SM	SSE	ESE	SMSE	SM	SSE	ESE	SMSE
0.2	Age	$\hat{\beta}_1$	0.004	0.006	0.004	0.000	0.000	0.003	0.003	0.000
	Male Vs Female	$\hat{\beta}_2$	0.612	0.097	0.081	0.384	0.001	0.094	0.086	0.009
	Smoker Vs Non	$\hat{\beta}_3$	0.160	0.102	0.083	0.036	-0.003	0.086	0.082	0.007
	$L_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_1$					0.995	0.096	0.091	0.009
	$L_I H_E$ Vs $H_I H_E$	$\hat{\gamma}_2$					0.992	0.117	0.110	0.014
	$M_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_3$					0.497	0.095	0.090	0.009
	$M_I H_E$ Vs $H_I H_E$	$\hat{\gamma}_4$					0.488	0.137	0.128	0.019
	$H_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_5$					0.197	0.288	0.231	0.086
		$\hat{\rho}_{WMM}$	0.423	0.073		0.055	0.158	0.081		0.008
0.4	Age	$\hat{\beta}_1$	0.004	0.007	0.005	0.000	0.000	0.004	0.004	0.000
	Male Vs Female	$\hat{\beta}_2$	0.624	0.098	0.086	0.399	-0.001	0.107	0.100	0.011
	Smoker Vs Non	$\hat{\beta}_3$	0.106	0.087	0.077	0.019	0.000	0.084	0.079	0.007
	$L_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_1$					0.989	0.113	0.106	0.013
	$L_I H_E$ Vs $H_I H_E$	$\hat{\gamma}_2$					0.989	0.138	0.126	0.019
	$M_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_3$					0.504	0.110	0.102	0.012
	$M_I H_E$ Vs $H_I H_E$	$\hat{\gamma}_4$					0.483	0.155	0.146	0.024
	$H_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_5$					0.226	0.306	0.262	0.094
		$\hat{\rho}_{WMM}$	0.552	0.070		0.028	0.315	0.088		0.015

(Table 5.1 contd...)

			Sampling Scheme							
			SRS				StRS			
ρ	Covariate	Parameter	SM	SSE	ESE	SMSE	SM	SSE	ESE	SMSE
0.6	Age	$\hat{\beta}_1$	0.004	0.007	0.005	0.000	0.000	0.005	0.004	0.000
	Male Vs Female	$\hat{\beta}_2$	0.631	0.113	0.091	0.411	0.005	0.126	0.117	0.016
	Smoker Vs Non	$\hat{\beta}_3$	0.072	0.084	0.068	0.012	-0.003	0.076	0.074	0.006
	$L_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_1$					0.989	0.133	0.123	0.018
	$L_I H_E$ Vs $H_I H_E$	$\hat{\gamma}_2$					0.981	0.160	0.147	0.026
	$M_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_3$					0.490	0.128	0.118	0.016
	$M_I H_E$ Vs $H_I H_E$	$\hat{\gamma}_4$					0.472	0.181	0.171	0.034
	$H_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_5$					0.175	0.383	0.312	0.153
		$\hat{\rho}_{WMM}$	0.667	0.061		0.008	0.474	0.088		0.024
0.8	Age	$\hat{\beta}_1$	0.004	0.008	0.005	0.000	0.000	0.006	0.005	0.000
	Male Vs Female	$\hat{\beta}_2$	0.636	0.126	0.096	0.420	0.006	0.155	0.138	0.024
	Smoker Vs Non	$\hat{\beta}_3$	0.048	0.078	0.058	0.008	0.000	0.077	0.068	0.006
	$L_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_1$					0.971	0.159	0.147	0.026
	$L_I H_E$ Vs $H_I H_E$	$\hat{\gamma}_2$					0.969	0.196	0.178	0.040
	$M_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_3$					0.485	0.148	0.139	0.022
	$M_I H_E$ Vs $H_I H_E$	$\hat{\gamma}_4$					0.473	0.215	0.197	0.047
	$H_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_5$					0.148	0.444	0.358	0.207
		$\hat{\rho}_{WMM}$	0.764	0.054		0.004	0.616	0.076		0.040

Table 5.2: Simulated sampling design based CWGQL (DBCWGQL) estimates for the diagnostic and/or prognostic covariates in MAR based incomplete longitudinal models for count data with non-response index parameter $\alpha = 5$, and selected values of the longitudinal correlation ρ (estimated by MM), based on 1000 simulations.

			Sampling Scheme							
			SRS				StRS			
ρ	Covariate	Parameter	SM	SSE	ESE	SMSE	SM	SSE	ESE	SMSE
0.2	Age	$\hat{\beta}_1$	0.004	0.006	0.004	0.000	0.000	0.004	0.003	0.000
	Male Vs Female	$\hat{\beta}_2$	0.620	0.097	0.083	0.394	0.000	0.094	0.087	0.009
	Smoker Vs Non	$\hat{\beta}_3$	0.143	0.101	0.081	0.031	-0.002	0.086	0.082	0.007
	L_IL_E Vs H_IH_E	$\hat{\gamma}_1$					0.995	0.095	0.092	0.009
	L_IH_E Vs H_IH_E	$\hat{\gamma}_2$					0.992	0.118	0.110	0.014
	M_IL_E Vs H_IH_E	$\hat{\gamma}_3$					0.494	0.096	0.090	0.009
	M_IH_E Vs H_IH_E	$\hat{\gamma}_4$					0.487	0.137	0.129	0.019
	H_IL_E Vs H_IH_E	$\hat{\gamma}_5$					0.195	0.292	0.232	0.088
		$\hat{\rho}_{MM}$	0.467	0.068		0.076	0.219	0.083		0.007
0.4	Age	$\hat{\beta}_1$	0.004	0.007	0.005	0.000	0.000	0.004	0.004	0.000
	Male Vs Female	$\hat{\beta}_2$	0.630	0.098	0.088	0.407	-0.001	0.107	0.100	0.011
	Smoker Vs Non	$\hat{\beta}_3$	0.094	0.085	0.075	0.016	0.000	0.083	0.078	0.007
	L_IL_E Vs H_IH_E	$\hat{\gamma}_1$					0.991	0.113	0.106	0.013
	L_IH_E Vs H_IH_E	$\hat{\gamma}_2$					0.991	0.138	0.127	0.019
	M_IL_E Vs H_IH_E	$\hat{\gamma}_3$					0.502	0.110	0.103	0.012
	M_IH_E Vs H_IH_E	$\hat{\gamma}_4$					0.482	0.155	0.146	0.024
	H_IL_E Vs H_IH_E	$\hat{\gamma}_5$					0.225	0.306	0.265	0.094
		$\hat{\rho}_{MM}$	0.588	0.065		0.040	0.374	0.085		0.008

(Table 5.2 contd...)

			Sampling Scheme							
			SRS				StRS			
ρ	Covariate	Parameter	SM	SSE	ESE	SMSE	SM	SSE	ESE	SMSE
0.6	Age	$\hat{\beta}_1$	0.004	0.007	0.005	0.000	0.000	0.005	0.004	0.000
	Male Vs Female	$\hat{\beta}_2$	0.636	0.112	0.092	0.417	0.005	0.126	0.117	0.016
	Smoker Vs Non	$\hat{\beta}_3$	0.064	0.081	0.066	0.011	-0.003	0.075	0.074	0.006
	$L_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_1$					0.990	0.133	0.124	0.018
	$L_I H_E$ Vs $H_I H_E$	$\hat{\gamma}_2$					0.983	0.160	0.148	0.026
	$M_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_3$					0.489	0.128	0.119	0.016
	$M_I H_E$ Vs $H_I H_E$	$\hat{\gamma}_4$					0.471	0.181	0.172	0.034
	$H_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_5$					0.176	0.382	0.313	0.151
		$\hat{\rho}_{MM}$	0.692	0.057		0.012	0.517	0.084		0.014
0.8	Age	$\hat{\beta}_1$	0.004	0.008	0.005	0.000	0.000	0.006	0.005	0.000
	Male Vs Female	$\hat{\beta}_2$	0.639	0.125	0.096	0.424	0.006	0.155	0.139	0.024
	Smoker Vs Non	$\hat{\beta}_3$	0.043	0.076	0.056	0.008	0.000	0.076	0.068	0.006
	$L_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_1$					0.972	0.159	0.147	0.026
	$L_I H_E$ Vs $H_I H_E$	$\hat{\gamma}_2$					0.970	0.196	0.179	0.039
	$M_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_3$					0.485	0.148	0.139	0.022
	$M_I H_E$ Vs $H_I H_E$	$\hat{\gamma}_4$					0.473	0.214	0.198	0.047
	$H_I L_E$ Vs $H_I H_E$	$\hat{\gamma}_5$					0.150	0.442	0.359	0.205
		$\hat{\rho}_{MM}$	0.776	0.052		0.003	0.638	0.071		0.031

Chapter 6

Concluding Remarks

To develop valid inference techniques in the incomplete longitudinal set-up for discrete data, it is important to understand both longitudinal correlation structure and the missing mechanism (such as MAR) involved in generating such data. As an improvement over the existing weighted generalized estimating equation approach, this thesis has developed a CWGQL (conditional weighted generalized quasilikelihood) approach by exploiting a conditional weighted distance function which accommodates both longitudinal correlation structure and the underlying missing mechanism. It is shown that this CWGQL produces regression estimates with much smaller bias than the existing approaches. The longitudinal correlation structures and the MAR missing mechanism are discussed in details in the thesis for both repeated binary and count data, the analysis of repeated incomplete count data being completely new.

The proposed CWGQL approach has also been applied to the complex survey based incomplete longitudinal data. It has been demonstrated that the use of a simpler sampling technique such as SRS (simple random sampling) may be detrimental in estimating the regression effects when it is known that the finite population may be of complex nature such as containing strata or clusters.

We remark that the ideas developed in the thesis should be extendable to various other situations where repeated missing data may follow higher order correlation structure and/or more complex missing mechanism such as nonignorable. Also, the repeated response indicators may be assumed to be correlated following a suitable correlation structure whereas we have assumed that they are independent conditional on the past responses. These and other similar generalizations are however beyond the scope of the present thesis.

Bibliography

[1] Amemiya, T. (1985) *Advanced Econometrics.* Harvard University Press, Cambridge, Massachusetts.

[2] Bahadur, R.R. (1961) A representation of the joint distribution of responses to n dichotomous items. In H. Solomon (Ed.). *Studies in item analysis and prediction* (pp. 158-168). Stanford: Stanford University Press.

[3] Binder, D.A. (1983) On the variance of asymptotically normal estimators from complex surveys. *International Statistitical Review*, **51**, 279-92.

[4] Campbell, J.T. (1934) The poisson correlation function. *Proceedings of the Edinburgh Mathematical Society*, **4**, 18-26.

[5] Cox, D.R. (1972) The analysis of multivariate binary data. *Applied Statistics*, **21**, 113-120.

[6] Dawss, M. and Teicher, H. (1957) On infinitely divisible random vectors. *Annals of Mathematical Statistics*, **28**, 461-470.

[7] Farrell, P.J. and Sutradhar, B.C. (2006) A non-linear conditional probability model for generating correlated binary data. *Statistics and probability Letters*, **76**, 353-61.

[8] Fitzmaurice, G.M., Laird, N.M. and Zahner, G.E.P. (1996) Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association*, **91**, 99-108.

[9] Holgate, P. (1964) Estimation for the bivariate Poisson distribution. *Biometrika*, **51**, 241-45.

[10] Johnson, N.L. and Kotz, S. (1969) *Discrete Distributions.* Houghton Mifflin, Boston.

[11] Kanter, M. (1975) Autoregression for discrete processes mod 2. *Journal of Applied Probability*, **12**, 371-75.

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

[13] Mallick, T.S. and Sutradhar, B.C. (2008) GQL versus conditional GQL inferences for non-stationary time series of counts with overdispersion. *Journal of Time Series Analysis*, **29**, 402-20.

[14] Manski, C.F. (1987) Semiparametric analysis of random effects linear models from binary panel data. *Econometrica*, **55**, 357-62.

[15] McCullagh, P., (1983) Quasi-likelihood functions. *Annals of Statistics*, **11**, 59-67.

[16] McKenzie, E. (1988) Some ARMA models for dependent sequences of Poisson counts. *Advances in Applied Probability*, **20**, 822-35.

[17] Paik, M.C. (1997) The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association*, **92**, 1320-29.

[18] Prentice, R.L. (1988) Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**, 1033-48.

[19] Qaqish, B.F. (2003) A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, **90**, 455-63.

[20] Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106-21.

[21] Rotnitzky, A., Robins, J.M. and Scharfstein, D.O. (1998) Semparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, **93**, 1321-39.

[22] Rubin, D.B. (1976) Inference and missing data. *Biometrika*, **63**, 581-92.

[23] Sutradhar, B.C. (2008) Inferences in familial Poisson mixed models for survey data. *Sankhya*, **70**, 18-33.

[24] Sutradhar, B.C. (2003) An overview on regression models for discrete longitudinal responses. *Statistical Science*, **18**, 377-93.

[25] Sutradhar, B.C. and Das, K. (1999) On the efficiency of regression estimators in generalized linear models for longitudinal data. *Biometrika*, **86**, 459-65.

[26] Sutradhar, B.C. and Farrell, P.J. (2007) On optimal lag 1 dependence estimation for dynamic binary models with application to Asthma data. *Sankhya*, **69**, 448-67.

[27] Sutradhar, B.C. Jowaheer, V. and Sneddon, G. (2008) On a unified generalized Quasi-likelihood approach for familial-longitudinal non-stationary count data. *Scandinavian Journal of Statistics*, **35**, 597-612.

[28] Sutradhar, B.C. and Kovacevic, M. (2000) Analysing ordinal longitudinal survey data: generalised estimating equations approach. *Biometrika*, **87**, 837-48.

[29] Teicher, H. (1954) On the multivariate Poisson distribution. *Scandinavisk Aktuarietidskrift*, **37**, 1-9.

[30] Wedderburn, R.W.M. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439-47.

[31] Williamson, J.M., Kim, K. and Lipsitz, S.R. (1995) Analyzing bivariate ordinal data using a global odds ratio. *Journal of the American Statistical Association*, **90**, 1432-37.

[32] Yi, G.Y. and Cook, R.J. (2002) Marginal methods for incomplete longitudinal data arising in clusters. *Journal of the American Statistical Association*, **97**, 1071-80.