ANALYSING LONGITUDINAL DATA IN THE PRESENCE OF MISSING RESPONSES WITH APPLICATION TO SLID DATA

ADEBOLA BRAIMOH

Analysing Longitudinal Data in the Presence of Missing Responses with Application to SLID Data

by

©Adebola Braimoh

A practicum submitted to the School of Graduate Studies
in partial fulfillment of the requirement for the Degree of
Master of Applied Statistics

Department of Mathematics and Statistics
Memorial University of Newfoundland

January 16, 2004

St. John's **Newfoundland and Labrador** **Canada**

Abstract

In longitudinal studies, outcomes that are repeatedly measured over time may be correlated and some may be missing. In this practicum, we empirically examine the performance of a recently proposed generalized quasi-likelihood (GQL) approach for the analysis of longitudinal data that includes observation that are missing completely at random (MCAR) or missing at random (MAR). This GQL approach is also illustrated by reanalyzing the Survey of Labour and Income Dynamics (SLID) data from Statistics Canada.

Acknowledgements

Firstly, I dedicate this work to the mercies of Almighty Allah.

I am grateful to my supervisor, Professor B.C. Sutradhar, for leading me through this field of research, and also for his relentless, helpful and thoughtful comments, discussions and suggestions throughout the preparation of this practicum. This work would not have been completed without the guidance, advice, encouragement, understanding and support of my supervisor during my programme. He has been generous with his ideas and time.

I am grateful to the School of Graduate Studies and the Department of Mathematics and Statistics for providing me with financial support in the form of Graduate Student Scholarships and Graduate Assistantships, and also providing me the opportunity to enhance my teaching experience during my study.

I would like to thank the Department, especially the statistics group for providing me with a very friendly atmosphere and facilities to complete my programme.

I am grateful to my parents, brothers and sisters for their love, support and encouragement throughout my university career.

It is my pleasure to thank all my friends and well-wishers who encouraged me during my programme.

Finally, I want to express my sincere appreciation to my wife, Aminat and my little daughter Aisha for their continued support, encouragement, care, understanding and love.

Contents

List of Tables

Chapter 1

Introduction

1.1 Motivation of the Problem

In many socio-economic research fields, it is common to collect observations successively over time on a large number of individuals. Also, a set of multidimensional covariates is often collected for each of these individuals. As the responses are collected repeatedly, it is likely that they will be correlated. In this type of longitudinal set up, it is of interest to find the effects of the covariates after taking the longitudinal correlations of the responses into account. This is, however, not easy as in practice the joint distribution of the correlated responses is not available.

Liang and Zeger (1986) have bypassed the joint distribution and used a 'working' correlation approach for the analysis of longitudinal data. This approach, however, has many pitfalls as shown by Crowder (1995) and Sutradhar and Das (1999). As a remedy, Sutradhar and Das (1999)[see also Jowaheer and Sutradhar (2002)] have suggested a true robust autocorrelation structure-based generalized quasilikelihood (GQL) approach to construct consistent as well as efficient regression estimates.

Note that in practice, it may happen that some of the repeated data collected over time may be missing for some individuals. The analysis of such longitudinal data subject to non-response is naturally more complicated. Some authors such as Paik (1997), Xie and Paik (1997), Robins, Rotnitzky, and Zhao (1995)(hereafter called

RRZ (1995))[see also Robins and Rotnitzky (1995)] have extended the 'working' correlation based generalized estimating equation (GEE) approach of Liang and Zeger (1986) to analyze such longitudinal data subject to non-response. More specifically, Paik (1997) has used the 'working' independence approach as a special case of the 'working' GEE approach. Note however that even though the independence approach may be efficient in some cases, it follows from Sutradhar and Das (1999) that it may be inefficient in some cases, specially when longitudinal data follow an AR(1) correlation structure. Consequently, these 'working' independence or general 'working' correlations based approaches run into difficulties in estimating the regression effects efficiently. Sutradhar and Kovacevic (2003) recently proposed an extension of the GQL approach of Sutradhar and Das (1999) and analyzed longitudinal data subject to non-response when responses occur either MCAR (missing completely at random) or MAR (missing at random). It is not known however, how the efficiency of such an extended GQL method will vary depending on the missingness structure especially when responses are MAR. This motivated us to undertake an empirical study to examine the efficiency of the GQL approach for various missing value structures.

1.2 Objective of the Practicum

As mentioned in the previous section, the main objective of this practicum is to examine the efficiencies of the GQL approach used by Sutradhar and Kovacevic (2003) in estimating the effects of the covariates in the longitudinal set up when the longitudinal response may be subject to non-response. We also apply the GQL approach for MAR models to the SLID (Survey of Labour and Income Dynamics) data collected by Statistics Canada for the period 1993 to 1998. The specific plan of the practicum is as follows.

In Chapter 2, we discuss the MCAR (Missing Completely at Random) and MAR (missing at random) models and summarize the GQL estimation approach for longitudinal data that follow either MCAR or MAR models. Note that both monotonic and non-monotonic missing cases are discussed.

In Chapter 3, we conduct a rigorous simulation study to examine the relative performance of the GQL approach under complete and various incomplete (subject to missing) longitudinal models. Once again, both monotonic and non-monotonic cases are considered.

In Chapter 4, we introduce the SLID data subject to non-response. We then apply the GQL methodology discussed in Chapters 2 and 3 to the SLID data.

We provide some concluding remarks in Chapter 5.

Chapter 2

Generalized Quasilikelihood Approach for Longitudinal Data Either MCAR or MAR

2.1 Background

In the longitudinal set up, a number of responses are collected repeatedly from a large number of individuals. Also, a set of covariates is collected from each of the individuals. Let

$$Y_i^c = (y_{i1}, \ldots, y_{it}, \ldots, y_{iT})' \; and \; X_i^c = (x_{i1}, \ldots, x_{it}, \ldots, x_{iT})' \tag{2.1}$$

denote the $T \times 1$ complete outcome vector and $T \times p$ covariate matrix respectively for the $ith (i = 1, \ldots, K)$ individual. Further let β be the effect of x_{it} on y_{it} for all $i = 1, \ldots, K$ and $t = 1, \ldots, T$. It is of interest to compute this β consistently and efficiently. Under the assumption that

$$E(Y_{it}) = a'(\theta_{it}) = \mu_{it} \; and \; Var(Y_{it}) = a''(\theta_{it}) \tag{2.2}$$

with $a(\theta_{it})$ as a known function of $\theta_{it} = x'_{it}\beta$ and $a'(\theta_{it})$ and $a''(\theta_{it})$ are the first and second derivatives of $a(\theta_{it})$ with respect to θ_{it}, one may obtain a consistent estimator of β by solving the so-called independence estimating equation

$$\sum_{i=1}^{K} X_i^{c'}(y_i^c - \mu_i^c) = 0 \tag{2.3}$$

where $X_i^{c'} = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, and $\mu_i^c = (\mu_{i1}, \ldots, \mu_{iT})'$. Note that to construct (2.3), it was assumed that $Var(Y_i^c) = A_i^c = \text{diag}[a''(\theta_{i1}), \ldots, a''(\theta_{it}), \ldots, a''(\theta_{iT})]$. Let $\hat{\beta}_I$ be the solution of (2.3), which is known to be consistent. As the repeated data $y_{i1}, \ldots, y_{it}, \ldots, y_{iT}$ are likely to be correlated, $\hat{\beta}_I$ obtained from (2.3) may not be efficient in all cases.

To obtain a consistent and efficient estimator, one may follow Jowaheer and Sutradhar (2002)[see also Sutradhar and Das(1999)] and solve the estimating equation

$$\sum_{i=1}^{K} X_i^{c'} A_i^c {\Sigma_i^c}^{-1}(y_i^c - \mu_i^c) = 0 \tag{2.4}$$

where $\Sigma_i^c = Var(Y_i^c) = A_i^{c^{1/2}} C(\rho) A_i^{c^{1/2}}$, with $C(\rho)$ as a $T \times T$ general auto-correlation matrix given by

$$C(\rho_1, \ldots, \rho_{T-1}) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{T-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_{T-1} & \rho_{T-2} & \rho_{T-3} & \cdots & 1 \end{bmatrix}, \tag{2.5}$$

ρ_ℓ being the ℓth lag autocorrelation which can be calculated as

$$\hat{\rho}_\ell = \frac{\sum_{i=1}^{K} \sum_{t=1}^{T-\ell} \tilde{y}_{it}\tilde{y}_{i,\,t+\ell}/K(T-\ell)}{\sum_{i}^{K} \sum_{t=1}^{T} \tilde{y}_{it}^2/KT} \tag{2.6}$$

with standardized residuals $\tilde{y}_{it} = (y_{it} - \mu_{it})/\{a''(\theta_{it})\}^{1/2}$

Let $\hat{\beta}_G$ be the solution of (2.4), which is consistent as well as highly efficient.

In the above discussion, it was assumed that all K individuals had responses for all T occasions. In practice, it may however happen that some of the repeated responses of an individual are missing.

Let r_{it} be an indicator variable, such that

$$r_{it} = \begin{cases} 1 & \text{if } y_{it} \text{ is observed} \\ 0 & \text{if } y_{it} \text{ is missing} \end{cases}$$

Suppose that for the ith individual $\sum_{t=1}^{T} r_{it} = T_i \leq T$.

Let $Y_i = (y_{i1}, \ldots, y_{it}, \ldots, y_{iT_i})'$ be the response vector for the ith individual, and $X_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}, \ldots, \mathbf{x}_{iT_i})'$ be the corresponding covariate matrix. Clearly $T - T_i$ responses are missing. Note that these $T - T_i$ missing responses may be monotonic or non-monotonic. To be specific, the missing responses will be monotonic if

$$r_{i1} \geq r_{i2} \geq \ldots \geq r_{it} \ldots \geq r_{iT_i} \ldots \geq r_{iT}\,, \tag{2.7}$$

otherwise the responses will be non-monotonically missing. For example, $r_{i1} = r_{i2} = r_{i3} = 1$, $r_{i4} = 0, r_{i5} = 1 \ldots r_{iT} = 0$ indicate that the missing responses are non-monotonic. Further note that if the data contain missing responses, then one cannot use the estimating equations (2.3) or (2.4) as they are constructed for the estimation of the parameters based on complete data. Some authors, such as Paik (1997) consequently modified the estimating equations (2.3) and (2.4) to obtain consistent estimates for the parameters based on incomplete data. These modifications however require the knowledge of missing patterns such as whether missingness is monotonic (such as in (2.7)) or not. Second, these modifications also require a probability model for $r_{it}(t = 1, \ldots, T)$ in order to determine the weights for respective y_{it}.

We now define certain non-response mechanisms (probability models for non-responses) that have been widely used in the literature. These non-response mechanisms are classified into three (3) categories (Little and Rubin,1987): MCAR (Missing Completely At Random), MAR (Missing At Random) and Non-Ignorable .

To elaborate, let $r_i = (r_{i1}, \ldots, r_{it}, \ldots, r_{iT_i})'$ be a vector of indicator variables for the ith subject, where as before, $r_{it} = 1$ if y_{it} is observed and $r_{it} = 0$ if y_{it} is missing.

Given r_i, the complete-data vector $\mathbf{Y}_i^c$ can be partitioned as $\mathbf{Y}_i^c = (\mathbf{Y}_{oi}, \mathbf{Y}_{mi})$, where $\mathbf{Y}_{oi}$ are the values of $\mathbf{Y}_i^c$ that are observed and $\mathbf{Y}_{mi}$ denotes the components of $\mathbf{Y}_i^c$ that are missing. Next let γ denote the vector of parameters of the nonresponse model so that $f(\mathbf{r}_i|\mathbf{y}_i^c, x_i^c, \gamma)$ denotes the joint distribution of $\mathbf{r}_i$ given $\mathbf{Y}_i^c$ and γ. In this notation the responses are *missing completely at random* (MCAR) if :

$$f(\mathbf{r}_i|\mathbf{y}_i^c, x_i^c, \gamma) = f(\mathbf{r}_i|x_i^c, \gamma) \tag{2.8}$$

(missingness does not depend on the values of the data $\mathbf{Y}_i^c$) and they are *missing at random* (MAR) if:

$$f(\mathbf{r}_i|\mathbf{y}_i^c, x_i^c, \gamma) = f(\mathbf{r}_i|\mathbf{y}_{oi}, x_i^c, \gamma) \tag{2.9}$$

(missingness depends only on the components $\mathbf{Y}_{oi}$ of $\mathbf{Y}_i^c$ that are observed, and not on the component that are missing). Finally, the missing data mechanism is *nonignorable* if:

$$f(\mathbf{r}_i|\mathbf{y}_i^c, x_i^c, \gamma) = f(\mathbf{r}_i|\mathbf{y}_{mi}, x_i^c, \gamma) \tag{2.10}$$

that is, the probability of nonresponse depends on the missing values $\mathbf{Y}_{mi}$, so that

$$f(\mathbf{y}_{oi}, \mathbf{r}_i|x_i, \gamma) = \sum_{Y_{m_i}} f(\mathbf{r}_i|\mathbf{y}_i, x_i, \gamma) f(\mathbf{y}_i|x_i),$$

where summation is over all possible values of $\mathbf{Y}_{m_i}$.

As examples, recently Paik (1997) has considered the following MAR and nonignorable mechanisms in a longitudinal study with monotonic missing responses.

M1: $Pr(r_{it} = 1|\mathbf{Y}_i^c, \mathbf{x}_i, r_{it-1} = 1) = Pr(r_{it} = 1|y_{i1}, \mathbf{x}_i, r_{it-1} = 1)$

M2: $Pr(r_{it} = 1|\mathbf{Y}_i^c, \mathbf{x}_i, r_{it-1} = 1) = Pr(r_{it} = 1|y_{i1}, \ldots, y_{it-1}, \mathbf{x}_i, r_{it-1} = 1)$, and

M3: $Pr(r_{it} = 1|\mathbf{Y}_i^c, \mathbf{x}_i, r_{it-1} = 1) = Pr(r_{it} = 1|y_{i1}, \ldots, y_{it}, \mathbf{x}_i, r_{it-1} = 1)$ respectively.

M1 and M2 are MAR (Rubin 1976), and M3 is nonignorable (Laird 1988; Little and Rubin 1987).

2.2 GQL Approach for Longitudinal Data MCAR

In this sub-section, we concentrate our discussion on the analysis of incomplete data when missing values occur completely at random. To be specific, the missingness does not depend on the data (see eq.(2.8)). This implies that $E\{r_{it}(Y_{it} - \mu_{it})\} = 0$ under this mechanism. Note that in practice, the missingness may occur monotonically or arbitrarily. In the next subsections, we deal with the estimating of the regression parameters for these two cases.

2.2.1 Monotonic Missing Case

For this type of longitudinal data MCAR, RRZ (1995) and Paik (1997, Section 2, p. 1320) suggest using the 'working' correlation matrix based estimating equation

$$U(\beta, \alpha) = \sum_{i=1}^{K} \frac{\partial \mu_i^{c'}}{\partial \beta} [\Sigma_i^c(\beta, \alpha)]^{-1} R_i (Y_i^c - \mu_i^c) = 0 \tag{2.11}$$

for the estimation of the regression parameter vector β , where $\Sigma_i^c = [A_i^c]^{1/2} R^*(\alpha) [A_i^c]^{1/2}$ with $A_i^c = \text{diag}[var(Y_{i1}), \ldots, var(Y_{iT})]$, and $R^*(\alpha)$ is a suitable $T \times T$ 'working' correlation matrix. Furthermore, in (2.11), $R_i = \text{diag}[r_{i1}, \ldots, r_{it}, \ldots, r_{iT}]$ with $r_{i1} \geq r_{i2} \geq \ldots \geq r_{it} \ldots \geq r_{iT_i} \ldots \geq r_{iT}$. This 'working' correlation matrix based approach has, however, many pitfalls. See Sutradhar and Das (1999) and Crowder (1995) with regard to this problem. In particular, this approach may produce inefficient estimates as compared to the 'working' independence approach. As a remedy, in order to obtain consistent and efficient estimator of β for the cases when longitudinal data are complete, Sutradhar and Kovacevic (2002) [see also Sutradhar and Das (1999)], have proposed a true correlation structure GQL structure based approach. This GQL approach based estimating equation for β is given by

$$U^*(\beta, \rho) = \sum_{i=1}^{K} \frac{\partial \mu_i^{c'}}{\partial \beta} [(I - R_i) + R_i \Sigma_i^c(\beta, \rho) R_i']^{-1} R_i (Y_i^c - \mu_i^c) = 0, \tag{2.12}$$

for the longitudinal responses MCAR with monotonic missing pattern. In (2.12) I

is the $T \times T$ identity matrix, and $\Sigma_i^c(\beta, \rho) = [A_i^c]^{1/2} C(\rho) [A_i^c]^{1/2}$, where $C(\rho)$ is the true correlation matrix of the data as defined in (2.5). Remark that unlike in RRZ (1995) and Paik (1997), we are now required to estimate this correlation matrix $C(\rho)$. In estimating the longitudinal correlation matrix $C(\rho)$, we note that when the data contain missing values in a monotonic pattern, the observed data form clusters with unequal sizes. This unbalanced situation was accommodated in the construction of the GQL type estimating equations (2.12). To estimate the correlation under this unbalanced situation, we use a modified formula

$$\hat{\rho}_\ell = \frac{\sum_{i=1}^K \sum_{t=1}^{T-\ell} r_{it} r_{i,\, t+\ell} z_{it} z_{i,\, t+\ell} / \sum_{i=1}^K \sum_{t=1}^{T-\ell} r_{it} r_{i,\, t+\ell}}{\sum_{i=1}^K \sum_{t=1}^T r_{it} z_{it}^2 / \sum_{i=1}^K \sum_{t=1}^T r_{it}} \tag{2.13}$$

which reduces to the estimating formula (2.6) when the data is complete. As before, $r_{it} = 1$ or 0 , and $z_{it} = (y_{it} - \mu_{it})/\{var(Y_{it})\}^{1/2}$, y_{it} being observed or unobserved responses for $t = 2, \ldots, T_i \leq T$. Note that $\hat{\rho}_\ell$ computed by (2.13) is consistent for ρ_ℓ provided $\sum_{i=1}^K r_{iT}$ is reasonably large. This is because if $\sum_{i=1}^K r_{iT}$ is large, $\sum_{i=1}^K r_{it}$ for $t = 1, \ldots, T-1$, for example, would be much larger because of the monotonic missing pattern, leading to the consistency of $\hat{\rho}_\ell$ for all $\ell = 1, \ldots, T-1$.

Once $\hat{\rho}_\ell$ is computed by (2.13), these are used in (2.12) to obtain the estimate of the regression parameter vector β. The solution of (2.12), denoted by $\hat{\beta}_{GQL,MCAR}$, is obtained iteratively by using the iterative equation

$$\begin{aligned} \hat{\beta}_{GQL,MCAR}(m+1) &= \hat{\beta}_{GQL,MCAR}(m) + \left[\sum_{i=1}^K \frac{\partial \mu_i^{c'}}{\partial \beta} \{(I - R_i) + R_i \Sigma_i^c(\beta, \hat{\rho}) R_i'\}^{-1} R_i \frac{\partial \mu_i^c}{\partial \beta'} \right]_m^{-1} \\ &\quad \times \left[\sum_{i=1}^k \frac{\partial \mu_i^{c'}}{\partial \beta} \{(I - R_i) + R_i \Sigma_i^c(\beta, \hat{\rho}) R_i'\}^{-1} R_i (Y_i^c - \mu_i^c) \right]_m \end{aligned} \tag{2.14}$$

where $[.]_m$ denotes that the expression within the brackets is evaluated at $\hat{\beta}_{GQL,MCAR}(m)$, the value of $\hat{\beta}_{GQL,MCAR}$ at the mth iteration. Under some mild conditions, $\hat{\beta}_{GQL,MCAR}$ is asymptotically distributed as normal with mean β and covariance matrix, $cov(\hat{\beta}_{GQL,MCAR})$, given by

$$Cov(\hat{\beta}_{GQL,MCAR}) = \left[\sum_{i=1}^{K} \frac{\partial \mu_i^{c'}}{\partial \beta} \{(I - R_i) + R_i \Sigma_i^c(\beta, \hat{\rho}) R_i'\}^{-1} R_i \frac{\partial \mu_i^c}{\partial \beta'}\right]^{-1}$$

$$\times \left[\sum_{i=1}^{K} \frac{\partial \mu_i^{c'}}{\partial \beta} \{(I - R_i) + R_i \Sigma_i^c(\beta, \hat{\rho}) R_i'\}^{-1} R_i \Sigma_i R_i' \{(I - R_i) + R_i \Sigma_i^c(\beta, \hat{\rho}) R_i'\}^{-1} \frac{\partial \mu_i^c}{\partial \beta'}\right]$$

$$\times \left[\sum_{i=1}^{K} \frac{\partial \mu_i^{c'}}{\partial \beta} \{(I - R_i) + R_i \Sigma_i^c(\beta, \hat{\rho}) R_i'\}^{-1} R_i \frac{\partial \mu_i^c}{\partial \beta'}\right]^{-1} \quad (2.15)$$

2.2.2 Non Monotonic Missing Case

Using the indicator variable r_{it}, a matrix R_i is generated first, reflecting the present non-monotonic pattern. For example, consider a longitudinal case with T= 4. Suppose that for the ith individual, a response was missing at time t= 3 ($r_{i1} = 1$, $r_{i2} = 1$, $r_{i3} = 0$, and $r_{i4} = 1$). We then generate the R_i matrix for the ith individual following these non-monotonic ($r_{i1} = r_{i2} > r_{i3} < r_{i4}$) pattern. That is $R_i = \text{diag}[1, 1, 0, 1]$. Next, for the sake of using this information in an estimating equation, we construct a new but monotonic type response indicator matrix $\tilde{R}_i$ defined as $\tilde{R}_i = \text{diag}[1, 1, 1, 0]$. Note that because of this change, the position of the 3rd and 4th responses in the longitudinal sequence have been interchanged. To make it much clearer , the non-response positions indicated by 0 in the R_i matrix are shifted to the end in the new sequence i.e forming the $\tilde{R}_i$ matrix.

We now construct new correlation and covariance matrices following the above 'shifting' technique. Recall that $C(\rho)$ and Σ_i^c are the original correlation and covariance matrices , whereas we will refer to the new 'shifting' matrices by $\tilde{C}(\rho)$ and $\tilde{\Sigma}_i^c$ respectively.

To be specific, rewrite $C(\rho)$ and Σ_i^c matrices as follows for T=4 ;

$$C(\rho) = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{pmatrix}$$

and

$$\Sigma_i^c = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix}.$$

As the $\tilde{R}_i$ matrix was constructed by bringing the non-missing responses together (at the beginning of the sequence), we reflect this shifting on the above correlation and covariance matrices by bringing together the rows and columns of these matrices corresponding to the non-missing responses. That is , the rows and columns of these matrices corresponding to the missing responses are shifted to the end. Suppose that the new matrices are denoted by $C^*(\rho)$ and $\Sigma_i^{*^c}$ respectively. For the above example, these matrices are constructed as

$$C^*(\rho) = \begin{pmatrix} 1 & \rho_1 & \rho_3 & \rho_2 \\ \rho_1 & 1 & \rho_2 & \rho_1 \\ \rho_3 & \rho_2 & 1 & \rho_1 \\ \rho_2 & \rho_1 & \rho_1 & 1 \end{pmatrix}$$

and

$$\Sigma_i^{*^c} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{14} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{24} & \sigma_{23} \\ \sigma_{41} & \sigma_{42} & \sigma_{44} & \sigma_{43} \\ \sigma_{31} & \sigma_{32} & \sigma_{34} & \sigma_{33} \end{pmatrix}$$

following the position of responses and nonresponses in the $\tilde{R}_i$. As it is impossible (without imputation) to calculate correlations corresponding to the missing values, without any loss of generality, we can put zero in the last column and last row of $C^*(\rho)$ and $\Sigma_i^{*^c}$ matrices, as these rows and columns reflect the missing responses. Thus, we construct the final correlation and covariance matrices as

$$\tilde{C}(\rho) = \begin{pmatrix} 1 & \rho_1 & \rho_3 & 0 \\ \rho_1 & 1 & \rho_2 & 0 \\ \rho_3 & \rho_2 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$\tilde{\Sigma}_i^c = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{14} & 0 \\ \sigma_{21} & \sigma_{22} & \sigma_{24} & 0 \\ \sigma_{41} & \sigma_{42} & \sigma_{44} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

where $\tilde{\Sigma}_i^c$ can be calculated using $\tilde{\Sigma}_i^c = [A_i^c]^{1/2}\tilde{C}(\rho)[A_i^c]^{1/2}$ provided X_i is known. Consequently, for the non-monotonic case, the GQL approach based estimating equation for β is now given by

$$\tilde{U}(\beta, \rho) = \sum_{i=1}^{K} \frac{\partial \mu_i^{c'}}{\partial \beta}[(I - \tilde{R}_i) + \tilde{R}_i\tilde{\Sigma}_i^c(\beta, \rho)\tilde{R}_i']^{-1}\tilde{R}_i(Y_i^c - \mu_i^c) = 0, \tag{2.16}$$

for the longitudinal responses MCAR with nonmonotonic missing pattern.

For the computation of the $\tilde{C}(\rho)$ matrix involved in the $\tilde{\Sigma}_i^c$ matrix , we can still use the lag correlation estimating equation (2.13). Once $\hat{\rho}_\ell$ is computed by (2.13), these are used in (2.16) to obtain the estimate of the regression parameter vector β . The solution of (2.16), denoted by $\hat{\beta}_{GQL,MCAR}$ is obtained iteratively by using the equation

$$\begin{aligned} \hat{\beta}_{GQL,MCAR}(m+1) &= \hat{\beta}_{GQL,MCAR}(m) + \left[\sum_{i=1}^{K} \frac{\partial \mu_i^{c'}}{\partial \beta}\{(I - \tilde{R}_i) + \tilde{R}_i\tilde{\Sigma}_i^c(\beta, \hat{\rho})\tilde{R}_i'\}^{-1}\tilde{R}_i\frac{\partial \mu_i^c}{\partial \beta'}\right]_m^{-1} \\ &\times \left[\sum_{i=1}^{k} \frac{\partial \mu_i^{c'}}{\partial \beta}\{(I - \tilde{R}_i) + \tilde{R}_i\tilde{\Sigma}_i^c(\beta, \hat{\rho})\tilde{R}_i'\}^{-1}\tilde{R}_i(Y_i^c - \mu_i^c)\right]_m \end{aligned} \tag{2.17}$$

where $[.]_m$ denotes that the expression within the brackets is evaluated at $\hat{\beta}_{GQL,MCAR}(m)$, the value of $\hat{\beta}_{GQL,MCAR}$ at the mth iteration. Under some mild conditions, $\hat{\beta}_{GQL,MCAR}$ is asymptotically distributed as normal with mean β and covariance matrix, $cov(\hat{\beta}_{GQL,MCAR})$ given by

$$cov(\hat{\beta}_{GQL,MCAR}) = \left[\sum_{i=1}^{K} \frac{\partial \mu_i^{c'}}{\partial \beta}\{(I-\tilde{R}_i)+\tilde{R}_i\tilde{\Sigma}_i^c(\beta,\hat{\rho})\tilde{R}_i'\}^{-1}\tilde{R}_i\frac{\partial \mu_i^c}{\partial \beta'}\right]^{-1}$$
$$\times \left[\sum_{i=1}^{K} \frac{\partial \mu_i^{c'}}{\partial \beta}\{(I-\tilde{R}_i)+\tilde{R}_i\tilde{\Sigma}_i^c(\beta,\hat{\rho})\tilde{R}_i'\}^{-1}\tilde{R}_i\Sigma_i\tilde{R}_i'\{(I-\tilde{R}_i)+\tilde{R}_i\tilde{\Sigma}_i^c(\beta,\hat{\rho})\tilde{R}_i'\}^{-1}\frac{\partial \mu_i^c}{\partial \beta'}\right]$$
$$\times \left[\sum_{i=1}^{K} \frac{\partial \mu_i^{c'}}{\partial \beta}\{(I-\tilde{R}_i)+\tilde{R}_i\tilde{\Sigma}_i^c(\beta,\hat{\rho})\tilde{R}_i'\}^{-1}\tilde{R}_i\frac{\partial \mu_i^c}{\partial \beta'}\right]^{-1} \quad (2.18)$$

2.3 GQL approach for Longitudinal Data MAR

Recall from (2.9) that if the data are MAR, then the probability of missingness, that is the probability of r_{it} depends on the past outcomes $y_{i1},\ldots,y_{i,t-1}$. Under this scenario, $E\{r_{it}(Y_{it}-\mu_{it})\} \neq 0$, and the root of the GQL (2.12)

$$U^*(\beta,\rho) = \sum_{i=1}^{K} \frac{\partial \mu_i^{c'}}{\partial \beta}[(I-R_i)+R_i\Sigma_i^c(\beta,\rho)R_i']^{-1}R_i(Y_i^c-\mu_i^c) = 0, \quad (2.19)$$

is a biased estimate of β. To remove this bias, RRZ (1995, Section 3, p.109) proposed a weighted generalized estimating equation (WGEE) approach which is a modification of the GEE given in (2.11). To be specific, RRZ's (1995) WGEE is given by

$$U(\beta,\alpha) = \sum_{i=1}^{K} \frac{\partial \mu_i^{c'}}{\partial \beta}[\Sigma_i^c(\beta,\alpha)]^{-1}\Delta_i(Y_i^c-\mu_i^c) = 0 \quad (2.20)$$

for obtaining unbiased GEE estimates under MAR, where $\Delta_i = \text{diag}[\delta_{i1},\ldots,\delta_{it},\ldots,\delta_{iT}]$, with

$$\delta_{it} = r_{it}/Pr\{(\prod_{j=1}^{t} r_{ij}) = 1|H_{i,t-1}, \gamma\} \tag{2.21}$$

H_{it} being the history of the data for the ith individual up to time t; that is, $H_{it} = (X_i, y_{i1}, \ldots, y_{it})$, and γ is, a q-dimensional (say) vector of additional parameters used to model the conditional mean relationship of r_{it} as a function of $y_{i1}, \ldots, y_{i,t-1}$. RRZ (1995) showed that if Δ_i is estimated consistently, the root of WGEE (2.20) is consistent and asymptotically normal under MAR and monotonic missing patterns. Remark that similar to Liang and Zeger (1986), as the 'working' covariance matrix $\Sigma_i^c(\beta, \alpha)$ in (2.20) is chosen by the investigator, this WGEE approach also has the same efficiency related pitfalls as that of the original GEE approach (Sutradhar and Das (1999)).

Now to obtain a consistent and efficient estimator of β for the case when longitudinal data are monotonically MAR, one may modify the GQL estimating equation (2.19) for the MCAR data, and write a WGQL estimating equation given by

$$U^*(\beta, \rho, \gamma) = \sum_{i=1}^{K} \frac{\partial \mu_i^{c'}}{\partial \beta} \Delta_i' \{(\Delta_i^* - \Delta_i) + \Delta_i \Sigma_i^c(\beta, \rho, \gamma) \Delta_i'\}^{-1} \Delta_i (Y_i^c - \mu_i^c) = 0 \tag{2.22}$$

where $\Delta_i^* = \text{diag}[\delta_{i1}, \ldots, \delta_{it_i}, 1, \ldots, 1]$ is a $T \times T$ diagonal matrix with the first t_i diagonal elements same as the non-zero t_i diagonal elements of the Δ_i matrix, and the remaining $T - t_i$ diagonal elements are 1.

For modelling δ_{it} in (2.21), that is, the non-zero diagonal elements of the Δ_i matrix, we refer to RRZ (1995) and Paik (1997) among others. More specifically to compute δ_{it} by (2.22), one is required to model $\bar{\lambda}_{it} = Pr(r_{it} = 1|r_{i(t-1)} = 1, H_{it})$. RRZ (1995, eqn (8),(9)), for example, have modelled this non-response probability as

$$\bar{\lambda}_{it}(\gamma) = Pr(r_{it} = 1|r_{i(t-1)} = 1, H_{it}) = \frac{e^{\gamma' h(w_{it})}}{1 + e^{\gamma' h(w_{it})}} \tag{2.23}$$

where γ is a $q \times 1$ vector of unknown parameter as mentioned before, and $h(w_{it})$ is a known function of w_{it} with $w_{it} \equiv (X_i, y_{i1}, \ldots, y_{i,\,t-1})$ explained by H_{it}. Note that γ in

(2.23) may be estimated by using the partial maximum likelihood (ML) estimation method (RRZ (1995)). Let $\hat{\gamma}$ be the partial Maximum Likelihood estimator (MLE) of γ. To be specific, $\hat{\gamma}$ maximizes the partial likelihood

$$\begin{aligned} L(\gamma) &= \prod_i L_i(\gamma) \\ &= \prod_i \prod_t \left[\{\bar{\lambda}_{it}(\gamma)\}^{r_{it}} \{1 - \bar{\lambda}_{it}(\gamma)\}^{1-r_{it}} \right]^{r_{i(t-1)}} \end{aligned} \tag{2.24}$$

Once γ is estimated by $\hat{\gamma}$, we calculate δ_{it} as $\delta_{it} = \frac{r_{it}}{\prod_{j=1}^{t} \bar{\lambda}_{ij}(\hat{\gamma})}$.

Similar to RRZ (1995), Paik (1997) has also used a model for the MAR nonresponse mechanism. For example, in Paik's simulation study, a logistic MAR model, namely, logit $\{Pr(r_{i2} = 1 | r_{i1} = 1, y_{i1})\} = y_{i1}$ was used.

In our simulation study in chapter 3, we will consider two MAR mechanisms. In one, the missingness probability will depend on the outcome obtained at time point $t = 1$, for the longitudinal case with $T = 4$. In the other mechanism, the missingness at time t will depend on the outcomes obtained at time points $t = 1$ and $t = 2$, for the case with $T = 4$. These two models are denoted by

M1 logit $\{Pr(r_{it} = 1)\} = y_{i1}$ with $r_{i1} = r_{i2} = 1$, and $t = 3, 4$

M2 logit $\{Pr(r_{it} = 1)\} = \gamma_1 y_{i1} + \gamma_2 y_{i2}$ with $r_{i1} = r_{i2} = 1$, and $t = 3, 4$

Given that Δ_i is known, we may solve the WGEE (2.20) for β by using the iterative equation

$$\begin{aligned} \hat{\beta}_{GQL,MAR}(m+1) = \hat{\beta}_{GQL,MAR}(m) + &\left[\sum_{i=1}^{K} \frac{\partial \mu_i^{c'}}{\partial \beta} \Delta_i' \{(\Delta_i^* - \Delta_i) + \Delta_i \Sigma_i^c(\beta, \rho, \gamma) \Delta_i'\}^{-1} \Delta_i \frac{\partial \mu_i^c}{\partial \beta'} \right]_m^{-1} \\ &\times \left[\sum_{i=1}^{K} \frac{\partial \mu_i^{c'}}{\partial \beta} \Delta_i' \{(\Delta_i^* - \Delta_i) + \Delta_i \Sigma_i^c(\beta, \rho, \gamma) \Delta_i'\}^{-1} \Delta_i (Y_i^c - \mu_i^c) \right]_m \end{aligned} \tag{2.25}$$

where $[.]_m$ denotes that the expression within the brackets is evaluated at $\hat{\beta}_{GQL,MAR}(m)$, the value of $\hat{\beta}_{GQL,MAR}$ at the mth iteration. Note that the computation of the $\Sigma_i^c(\beta, \rho, \gamma)$ requires the estimation of $\rho = (\rho_1, \ldots, \rho_l, \ldots, \rho_{T-1})'$, which we obtain as

$$\hat{\rho}_\ell = \frac{\sum_{i=1}^{K}\sum_{t=1}^{T-\ell}\delta_{it}\delta_{i,\,t+\ell}z_{it}z_{i,\,t+\ell}/\sum_{i=1}^{K}\sum_{t=1}^{T-\ell}\delta_{it}\delta_{i,\,t+\ell}}{\sum_{i=1}^{K}\sum_{t=1}^{T}\delta_{it}z_{it}^2/\sum_{i=1}^{K}\sum_{t=1}^{T}\delta_{it}}, \tag{2.26}$$

following (2.13). In (2.26), $z_{it} = (y_{it} - \mu_{it})/\{var(Y_{it})\}^{1/2}$, y_{it} being observed or unobserved responses for $t = 2, \ldots, T_i \leq T$. Let $\hat{\beta}_{GQL,MAR}$ denote the WGEE based estimator of β obtained by (2.25). Under some mild conditions, it may be shown that $\hat{\beta}_{GQL,MAR}$ is asymptotically distributed as normal with mean β and covariance matrix, $cov(\hat{\beta}_{GQL,MAR})$, given by

$$\begin{aligned}
Cov(\hat{\beta}_{GQL,MAR}) \;=\; & \left[\sum_{i=1}^{K}\frac{\partial\mu_i^{c'}}{\partial\beta}\Delta_i'\{(\Delta_i^* - \Delta_i) + \Delta_i\Sigma_i^c(\beta,\rho,\gamma)\Delta_i'\}^{-1}\Delta_i\frac{\partial\mu_i^c}{\partial\beta'}\right]^{-1} \\
& \times\left[\sum_{i=1}^{K}\frac{\partial\mu_i^{c'}}{\partial\beta}\Delta_i'\{(\Delta_i^* - \Delta_i) + \Delta_i\Sigma_i^c(\beta,\rho,\gamma)\Delta_i'\}^{-1}\Delta_i\Sigma_i\Delta_i'\{(\Delta_i^* - \Delta_i)\right. \\
& \left. +\Delta_i\Sigma_i^c(\beta,\rho,\gamma)\Delta_i'\}^{-1}\frac{\partial\mu_i^c}{\partial\beta}\right] \\
& \times\left[\sum_{i=1}^{K}\frac{\partial\mu_i^{c'}}{\partial\beta}\Delta_i'\{(\Delta_i^* - \Delta_i) + \Delta_i\Sigma_i^c(\beta,\rho,\gamma)\Delta_i'\}^{-1}\Delta_i\frac{\partial\mu_i^c}{\partial\beta'}\right]^{-1}
\end{aligned} \tag{2.27}$$

Chapter 3

Performance Of the GQL Approach Under Complete and Various Incomplete Longitudinal Models: A Simulation Study

Recall that to analyze missing longitudinal data, Paik (1997) proposed to use a 'working' independence approach. But, as Sutradhar and Das (1999) have shown in the context of complete longitudinal data analysis that the 'working' independence approach may not be uniformly efficient as compared to the GEE approach discussed in the previous chapter. This mainly happens when the longitudinal data follow an AR(1) model. Further, to obtain a uniformly more efficient estimator for the regression parameter, Sutradhar and Das (1999) suggested the GQL approach (also discussed in the previous chapter) where the correlation structure is assumed to be known. In this chapter, we examine the performance of the GQL approach of Sutradhar and Das (1999), first for the complete AR(1) longitudinal data. More specifically in Section 3.1, we examine the performance of the GQL approach as compared to the 'working' independence approach in estimating β. This GQL method definitely appears to perform better as compared to the 'working' independence approach, as it appears to give nearly unbiased estimates of β with smallest mean square error. As the GQL approach performs better than GEE approach, we continue to examine its performance for the case when longitudinal data may be subject to non-response. In Section 3.2, we examine the performance of the GQL approach when the longitudinal data are MCAR. This is done monotonically in estimating β for various values of non-missing probabilities (NMP) for a response. Both monotonic and non-monotonic missing cases are discussed under MCAR models. By the same token, in Section 3.3, we examine the performance of the GQL approach when the longitudinal data are MAR. Here also we have incorporated both monotonic and non-monotonic missing cases. Note that the result of the simulation study presented in Section 3.1-3.3 should reveal the loss of efficiency because of missingness.

3.1 Performance Of the GQL Approach for Complete Longitudinal Data Analysis: Efficiency Comparison Between GQL and GEE(I) Approaches

Recall from Section 2.1 that in the GQL approach, one solves the estimating equation

$$\sum_{i=1}^{K} X_i^{c'} A_i^c \Sigma_i^{c^{-1}} (y_i^c - \mu_i^c) = 0 \tag{3.1}$$

for the regression parameter β. Let $\hat{\beta}_G$ denote this estimate.

This estimator with regard to the formula for covariance of $\hat{\beta}_G$, is consistent as the left hand side of (3.1) is an unbiased function of zero. Furthermore, this $\hat{\beta}_G$ is highly efficient as the estimating equation (3.1), similar to the traditional quasilikelihood approach, uses the true mean vector μ_i^c and the true covariance matrix Σ_i^c to construct the estimating equation. Nevertheless, Paik (1997) has used a 'working' independence assumption based GEE approach. In this section, we conducted a simulation study to compare the performance of the independence based GEE approach as compared to the GQL approach proposed by Sutradhar and Das (1999). For this purpose, we need to derive the formulas for this covariance of $\hat{\beta}_G$ and $\hat{\beta}_I$, where $\hat{\beta}_I$ is the 'working' independence based GEE estimator for β. With regard to the formula for the covariance of $\hat{\beta}_G$, it can be shown that under mild regularity conditions, $\hat{\beta}_G$ has the asymptotically covariance matrix given by

$$\left(\sum_{i=1}^{K} X_i^{c'} A_i^c \Sigma_i^{c^{-1}} A_i^c X_i^c\right)^{-1}. \tag{3.2}$$

If we, however, use the 'working' independence approach to estimate β, we then obtain the asymptotic covariance matrix of the estimator $\hat{\beta}_I$, given by

$$\left(\sum_{i=1}^{K} X_i^{c'} A_i^c X_i^c\right)^{-1} \left(\sum_{i=1}^{K} X_i^{c'} \Sigma_i^c X_i^c\right) \left(\sum_{i=1}^{K} X_i^{c'} A_i^c X_i^c\right)^{-1}. \tag{3.3}$$

This is because under the independence assumption, the estimating equation (3.1) reduces to

$$\sum_{i=1}^{K} X_i^{c'} (y_i^c - \mu_i^c) = 0 \tag{3.4}$$

Now to compare the variances of $\hat{\beta}_G$ and $\hat{\beta}_I$, we conduct a simulation study as follows: We consider $K = 100$ individuals and obtain $T = 6$ binary responses from each of the individuals following an AR(1) scheme. More specifically, to generate $y_{i1}, \ldots, y_{it}, \ldots, y_{iT}$, we follow a stationary AR(1) scheme for binary data as follows:

1. Generate binary y_{i1} with probability $\mu_{i.}$, where $\mu_{i.} = \frac{e^{x'_{i.}\beta}}{1+e^{x'_{i.}}}$

2. if y_{i1}=0, then generate binary y_{i2} with probability $\mu_{i.}(1-\rho)$; if y_{i1}=1, then generate y_{i2} with probability $\mu_{i.}$+ $\rho(1-\rho)$

3. Continue this to get y_{i3} depending only on y_{i2}, and so on.

   The above generating procedure ensures that the ℓth lag ℓ (ℓ=1,...,T-1) autocorrelation between y_{it} and $y_{i,t-\ell}$ is ρ^ℓ

As far as the covariates are concerned, we considered $p = 2$ time independent covariates. Also we follow two different designs D_1 and D_2 for this study. To be specific, under D_1 we consider

$$x_{it1} = \begin{cases} -1.0 & \text{for } i = 1, \frac{K}{4} \\ 0.0 & \text{for } i = \frac{K}{4} + 1, \frac{K}{2} \\ 0.0 & \text{for } i = \frac{K}{2} + 1, (\frac{K}{4}) * 3 \\ 1.0 & \text{for } i = (\frac{K}{4}) * 3 + 1, K \end{cases}$$

and $x_{it2} = z_i$, where $z_i (i = 1, \ldots, K)$ are generated independently from a normal distribution with mean zero and variance one. Under D_2, we consider the same x_{it1} but use $x_{it2} = t/6$ allowing certain time dependence. The component of β are denoted by β_1 and β_2 respectively. We consider the generation of the data for various large values of ρ, namely, ρ= 0.5, 0.8 and 0.9. Next, we compute $\hat{\beta}_G$ as a solution of $\sum_{i=1}^{K} X_i^{c'} A_i^c \Sigma_i^{c^{-1}} (y_i^c - \mu_i^c) = 0$ and $\hat{\beta}_I$ as a solution of $\sum_{i=1}^{K} X_i^{c'} (y_i^c - \mu_i^c) = 0$ respectively. This we do for 1000 simulations. We then compute the averages and standard errors of the 1000 simulated values of $\hat{\beta}_G$ and $\hat{\beta}_I$.

The simulated means (SM) and simulated standard errors (SSE) are reported in Table A.1. We also compute simulated mean square error (SMSE), where MSE= $(bias)^2 + SE^2$. This is also reported in Table 3.1. Further we compute the estimated standard error (ESE) by using the covariance matrices of $\hat{\beta}_G$ given in (3.2) and of $\hat{\beta}_I$ given in (3.3). We then take the average of the 1000 estimated standard errors and refer to them as ESE. These ESE are also reported in the same Table A.1. It is clear from Table A.1 that the GQL approach performs much better in estimating β parameters, as compared to the GEE(I) approach.

It is also clear from Table A.1 that the robust estimating formula (2.6) performs very well in estimating the correlation parameter. For example, for $T = 6$, ρ=0.8 under D_1, the estimates of $\rho_1, \ldots, \rho_5$ are found to be 0.7951, 0.6314, 0.5010, 0.3959 and 0.3155, which appears to agree with $\rho_\ell = \rho^\ell$ with ρ= 0.8, and $\ell = 1, \ldots, 5$. For the estimation of β, both $\hat{\beta}_G$ and $\hat{\beta}_I$ are unbiased. For example, for $T = 6$, ρ=0.8, the SMs of $\hat{\beta}_{1I}$ and $\hat{\beta}_{2I}$ are found to be 1.0263 and 1.0320 respectively, and the SMs of $\hat{\beta}_{1G}$ and $\hat{\beta}_{2G}$ are found to be 1.0284 and 1.0390 respectively. Thus, these regression estimates are unbiased with reference to the true value values of β, which are $\beta_1 = 1$ and $\beta_2 = 1$ respectively.

Next, the mean square error of $\hat{\beta}_G$ appears to be smaller than that of $\hat{\beta}_I$ irrespective of the values of T and ρ. The SMSE of $\hat{\beta}_G$ appears to be much smaller than that of $\hat{\beta}_I$ under design D_2. For example for $T = 6$, ρ=0.9 under D_1, the SMSE of $\hat{\beta}_{2G}$ is 0.2494, while the SMSE of $\hat{\beta}_{2I}$ is 0.9108. Thus, $\hat{\beta}_{2G}$ is 3.8 times more efficient than $\hat{\beta}_{2I}$.

In summary, $\hat{\beta}_G$ performs better in the sense that the SSE of $\hat{\beta}_G$ as well as the absolute values of their estimates of bias are smaller as compared to $\hat{\beta}_I$ for all times $T = 6$, 10 and 15 under both designs.

As in practice, one computes the estimated variance of the regression estimates, we have also computed the estimates of the variances of $\hat{\beta}_G$ and $\hat{\beta}_I$ by (3.2) and (3.3) respectively. These estimated standard errors are compared with the SSE given in the same Table 3.1.

The comparison of the SSE and ESE for $\hat{\beta}_G$ and $\hat{\beta}_I$ shows that the SSE for $\hat{\beta}_G$ is closer to its ESE, while the SSE of $\hat{\beta}_I$ is far away from its ESE. For example, using $T = 6$, ρ=0.9 under D_1, we have obtained the SSEs for $\hat{\beta}_{1I}$ and $\hat{\beta}_{2I}$ as 0.3378 and 0.3139 respectively, while the ESE are found to be 0.1841 and 0.1640 respectively. With regard to the performance of $\hat{\beta}_G$, we found the SSEs to be 0.3375 and 0.3092 for $\hat{\beta}_{1G}$ and $\hat{\beta}_{2G}$ respectively, while the ESE for $\hat{\beta}_{1G}$ and $\hat{\beta}_{2G}$ are found to be 0.3184 and 0.2864 respectively.

Thus, the GQL approach performs better as compared to GEE(I) in estimation. This result also holds for other time occasions $T = 10$ and 15.

3.2 Performance Of The GQL Approach For Longitudinal Data MCAR

Note that it is known from Sutradhar and Das (1999) that $\hat{\beta}_I$ is not uniformly better than the 'working' correlation based GEE estimators. In Section 3.1, as opposed to GEE, we have used the GQL approach suggested by Sutradhar and Das (1999, Section 3). It was demonstrated that when the data comes from an AR(1) process, for example, the GEE(I) performs worse as compared to the GQL approach. In view of these results, in this section, we use only the GQL approach and examine its performance for the cases when the data are MCAR.

We consider two cases for the MCAR model with various values of non-missing probabilities (NMP) for a response. Under the first case, we assume that data are monotonically missing completely at random, whereas, in the second case, data are assumed to be non-monotonically missing completely at random. The simulation studies for these cases are explained in Sections 3.2.1 and 3.2.2 respectively.

3.2.1 GQL Approach for Longitudinal Data Monotonically MCAR

Recall from Section 2.1 that when responses are missing in a monotonic pattern, then

$$r_{i1} \geq r_{i2} \geq \ldots \geq r_{it} \ldots \geq r_{iT_i} \ldots \geq r_{iT} \,, \tag{3.5}$$

where r_{it} is the response indicator for the tth $(t = 1, \ldots, T)$ observation of the ith $(i = 1, \ldots, K)$ individual. Now, in order to know whether r_{it}=1 or 0, one requires a probability model for $r_{it}(t = 1, \ldots, T)$. For convenience, we refer to $P(r_{it} = 1)$ as the non-missing probability (NMP). Under MCAR mechanism, this NMP does not depend on the outcomes. Consequently, one can consider an independent binary distribution for the selection of this probability. In the simulation study, we consider a wide variety of NMP such as $NMP \equiv 0.80, 0.90$ and 0.95.

Under MCAR, to generate $y_{it}(i = 1, \ldots, K; t = 1, \ldots, T_i)$, we first generate r_{it} for all $i = 1, \ldots, K$ and $t = 1, \ldots, T$ following the monotone pattern; that is once a subject leaves the study, return is not possible, or equivalently, r_{it}=0 implies that $r_{i(t+1)} = \ldots r_{iT}$=0. To be specific, under the present model, we consider $r_{i1} = r_{i2} = 1$ for all i. Now to generate r_{i3}, we generate this with binary probability NMP=$Pr(r_{i3} = 1)$=0.95, say. If $r_{i3} = 1$, we then generate r_{i4} with the same probability. If $r_{i3} = 0$, however, we put $r_{i3} = r_{i4} = \ldots = r_{iT} = 0$. Based on this MCAR mechanism, we now have the T_i ($T_i = 3, \ldots, T$) of non-missing response indicator for the ith individual. Consequently, we generate $y_{i1}, \ldots, y_{iT_i}$ for the ith individual following the procedure as in Section 3.1 for both design 1 (D_1) and design 2 (D_2). This means we use the covariate information x_{itu} ($t = 1, \ldots, T_i$; $u = 1, \ldots, p$) and the longitudinal correlation structure for generating $y_{i1}, \ldots, y_{iT_i}$ for all $i = 1, \ldots, K$.

For the regression parameters, we use the true value of β parameters, namely $\beta_1 = \beta_2 = 1$. Using the logistic model, we solve for the probabilities

$$p_{it} = \frac{exp\left[x_{it1}\beta_1 + x_{it2}\beta_2\right]}{1 + exp\left[x_{it1}\beta_1 + x_{it2}\beta_2\right]} \tag{3.6}$$

for $(i = 1, \ldots, K; t = 1, \ldots, T_i)$. Using the p_{it} results, we then compute the Σ_i, A_i and R_i matrices following their definition in Chapter 2, and estimate the regression parameter β and also compute the estimated covariance of the estimate of β, as in the previous chapter.

It is clear from Table A.2 that the correlation estimates obtained by using the unbalanced data under the present MCAR case appear to be highly satisfactory irrespective of the designs D_1 and D_2, values of T, and the non-missing probabilities. For example, under D_1, for T=6 and NMP=0.8, the correlation estimates are 0.89 ,0.79 ,0.70 , 0.63 and 0.56 whereas the true correlations, based on $\rho_\ell = \rho^\ell$, for ρ=0.9, are 0.9, 0.81, 0.72, 0.63 and 0.57 respectively. The standard errors of the correlation estimates appear to be reasonably small always. Similarly, for the same correlation parameter ρ=0.9, under D_2, and for T=10 and NMP=0.95, the first five correlation estimates are 0.90, 0.81, 0.73, 0.66 and 0.59 respectively, which are extremely close to the corresponding true values.

With regard to the estimation performance of the GQL approach in estimating β, it performs well in estimating both β_1 and β_2 under design 1 (D_1) especially when the NMP is large, as expected. For example, for T=6, NMP=0.95 and ρ=0.9, the simulated mean square errors (SMSEs) for β_1 and β_2 estimates are 0.12 and 0.11, whereas for NMP=0.8, the corresponding SMSEs are 0.16 and 0.13 respectively.

Similar results hold under D_2. This is because in general, the SMSEs are smaller when NMP is large. For example, for T=6, ρ=0.5 under D_2, the SMSEs for β_1 and β_2 estimates are 0.05 and 0.17 when NMP=0.95, but the SMSEs are 0.06 and 0.30 when NMP=0.80. We must also note that for a given NMP, say, NMP=0.95, the SMSEs for β_2 estimates under D_2 differs from that of D_1 considerably. This is because under D_2, the covariates were chosen to be time dependent i.e x_{it2}=t/T.

It is then clear that the GQL approach under MCAR model with high NMP does not perform well when the covariates are time dependent. This raises an issue of finding a better way to analyze the non-stationary missing data, which is, however, beyond the scope of the present practicum.

Remark that when the performance of the GQL approach under the MCAR model is compared with the non-missing case, the GQL performs better in the latter case, as expected. For example, when there were no missing values for the case with T=6, for ρ=0.9 under D_1, the SMSEs were found to be 0.12 and 0.10 (see Table 3.1), whereas for low NMP=0.8, the SMSEs for β_1 and β_2 estimates were 0.16 and 0.13 respectively.

3.2.2 GQL Approach For Longitudinal Data Non-Monotonically MCAR

Recall that when responses are non-monotonic in nature, for a given $t(t = 1, \ldots, T)$, r_{it} can take the value zero or one at random. Suppose that for an individual $i(i = 1, \ldots, K)$, the T responses are: $r_{i1} = r_{i2} = r_{i3} = 1$, $r_{i4} = 0, r_{i5} = 1 \ldots r_{iT} = 0$. Here, unlike in Section 3.2.1, r_{i5} can be 1 even if r_{i4}=0. This demonstrates an example for a non-monotonic missing case.

Under non-monotonic MCAR, to generate $y_{it}(i = 1, \ldots, K; t = 1, \ldots, T_i)$, we first consider $r_{i1} = r_{i2} = 1$ for all i. The remaining r_{it} for $t = 3, \ldots, T$ are generated randomly from a binary distribution with probability $P(r_{it} = 1)$=0.80, 0.90 and 0.95. Suppose that T_i values of $r_{it}(t = 1, \ldots, T)$ including for t=1 and t=2 are 1.

We now turn back to generate y_{it} corresponding to r_{it}=1. To do this, we first generate all y values, i.e, $y_{i1}, \ldots, y_{iT}$ following the AR(1) longitudinal correlation structure as discussed in Section 3.1. In the next step, we omit the values of y_{it} for which r_{it}=0. This generates y_{it} values in a non-monotonic fashion.

For the estimation of β by this GQL approach, we can follow the construction of the estimating equation as discussed in Section 2.2.2. The mean estimates also with the mean square error are reported in Table A.3 under both designs D_1 and D_2.

Unlike in the last section, here we consider T=4. This is because, unlike in the monotonic MCAR case, to construct the shifted covariance matrix gets complicated for large T. We however retain the same correlation values ρ=0.5, 0.8 and 0.9.

It is clear from Table A.3 that the correlation estimates appear to be highly satisfactory irrespective of the designs and the non-missing probabilities. For example, under D_1 with NMP=0.90, the correlation estimates are 0.89 , 0.80 , 0.72 whereas the true correlations based on $\rho_\ell = \rho^\ell$ for ρ=0.9 are 0.9, 0.81 and 0.72 respectively. The estimates approximately appear to satisfy the AR(1) relationship $\rho_\ell = \rho^\ell$. The standard errors of the correlation appear to be reasonably small always. Similarly, for the same correlation parameter ρ=0.9, under D_2 with NMP=0.95, the first three correlation estimates are 0.9, 0.81 and 0.72 respectively, which are extremely close to the corresponding true values.

The estimation of the parameter β appears to perform well under D_1, for both NMP=0.90 and 0.95. For example, for ρ=0.9, the simulated means of β_1 and β_2 are 1.045 and 1.062 for NMP=0.90 and 1.045 and 1.064 for NMP=0.95. The SMSE for these two cases are found to be 0.116 and 0.112 under NMP=0.90 and 0.117 and 0.106 under NMP=0.95. The performance of the GQL approach is, however, not quite satisfactory under the time dependent design D_2. For example, for the same parameter values under D_2, i.e, ρ=0.9, the SMSE for β_1 and β_2 are found to be 0.112 and 0.203 when NMP=0.90 and 0.096 and 0.205 when NMP=0.95. Thus, it is clear that the SMSE of $\hat{\beta}_2$ is much larger under D_2 than under D_1.

A comparison of the SSE and the ESE for the two non-missing probabilities shows that they are approximately close to each other. For example, using NMP=0.90, D_1 with ρ=0.8, the SSEs for $\hat{\beta}_1$ and $\hat{\beta}_2$ are found to be 0.25 and 0.23 respectively, while the ESEs are found to be 0.25 and 0.22 respectively. Likewise, for NMP=0.95, D_2 with ρ=0.8, the SSEs for $\hat{\beta}_1$ and $\hat{\beta}_2$ are found to be 0.30 and 0.31 respectively, while the ESEs are found to be 0.28 and 0.22 respectively.

Furthermore, a comparison of this non-monotonic MCAR with the monotonic MCAR shows that the estimates of the parameter under monotonic MCAR are approximately equal to the MCAR non-monotonic estimates under the two designs, with same ρ values. For example, for monotonic T=4, NMP=0.95, D_1 with ρ=0.8, the SMSEs for $\hat{\beta}_1$ and $\hat{\beta}_2$ are found to be 0.10 and 0.09 respectively, whereas, for the non-monotonic case T=4, NMP=0.95, D_1 and ρ=0.8, the SMSEs for $\hat{\beta}_1$ and $\hat{\beta}_2$ are estimated as 0.09 and 0.09 respectively. Similar results also hold under D_2. For example, for monotonic case T=4, NMP=0.95, D_2, ρ=0.8, the SMSEs for $\hat{\beta}_1$ and $\hat{\beta}_2$ are found to be 0.09 and 0.15 respectively, while for non-monotonic case T=4, NMP=0.95, D_2, ρ=0.8, the SMSEs for $\hat{\beta}_1$ and $\hat{\beta}_2$ are found to be 0.09 and 0.13 respectively. This shows that the efficiency performances of the GQL approach under monotonic and non-monotonic MCAR cases remain the same. The difference between the two approaches is that the estimation is more complicated under the non-monotonic case as compared to the monotonic case.

3.3 Performance Of The GQL Approach For Longitudinal Data MAR

Recall from (2.9) that if the data are MAR, then the probability of missingness, that is the probability of r_{it} depends on the past outcomes $y_{i1}, \ldots, y_{i,t-1}$, where r_{it} is the response indicator for the tth $(t = 1, \ldots, T)$ observation of the ith $(i = 1, \ldots, K)$ individual.

Now, in order to know whether r_{it}=1 or 0, one requires a probability model for $r_{it}(t = 1, \ldots, T)$. One can consider an independent binary distribution for the selection of this probability. We will illustrate this procedure for the cases when the MAR mechanism is monotonic and non-monotonic. As indicated in Section 2.3, the two MAR models **M1** and **M2** with probability logit $\{Pr(r_{it} = 1)\} = y_{i1}$ and logit $\{Pr(r_{it} = 1)\} = \gamma_1 y_{i1} + \gamma_2 y_{i2}$ respectively, will be considered in the simulation study. More specifically, we will consider γ_1 and γ_2 to have values 0.3 and 0.7 respectively. For **M1**, the missingness probability will depend on the outcome obtained at time point $t = 1$, and for **M2**, the missingness probability at time t will depend on the outcomes obtained at time $t = 1$ and $t = 2$.

3.3.1 Generation of the Data under MAR M1 and M2

As in the last section, we assume that $r_{i1} = r_{i2} = 1$. Consequently, y_{i1} and y_{i2} can be generated immediately. To be specific, we now generate y_{i1} and y_{i2} for the ith individual under D_1 and D_2 following the AR(1) scheme with correlation ρ as explained in Section 3.1.

Next, we generate r_{i3} with probability logit $\{Pr(r_{i3} = 1)\} = y_{i1}$ under **M1** and with logit $\{Pr(r_{i3} = 1)\} = \gamma_1 y_{i1} + \gamma_2 y_{i2}$ under **M2**. If r_{i3}=0, use $r_{i3} = r_{i4} = \ldots = r_{iT}$=0 under the monotonic approach. If however, r_{i3}=1, then generate y_{i3} following the AR(1) scheme by relating y_{i3} with y_{i2} and y_{i1} based on the mean and longitudinal correlation structures of the binary responses. Also, we continue to generate r_{i4} following **M1** and **M2** models. If r_{i4}=1 , we generate y_{i4} in the manner similar to the generation of y_{i3}. We continue this process until all $y_{it}(i = 1, \ldots, K; t = 1, \ldots, T_i)$ are generated for the ith individual.

We now explain how to generate the data in the non-monotonic case under models **M**1 and **M**2. Here, we first generate all $y_{it}(i = 1, \ldots, K; t = 1, \ldots, T)$ for the ith individual following the procedure as in Section 3.1 for both design 1 (D_1) and design 2 (D_2). All these y_{it}'s however will not be included in the data. In order to create a valid set of responses, we generate r_{it} for all $i = 1, \ldots, K$ and $t = 3, \ldots, T$ following the MAR models **M**1 and **M**2. That is, under **M**1 model, we generate $r_{it}(t = 3, \ldots, T)$ based on the probability model logit $\{Pr(r_{it} = 1)\} = y_{i1}$ and under model **M**2 with probability logit $\{Pr(r_{it} = 1)\} = \gamma_1 y_{i1} + \gamma_2 y_{i2}$. We now retain those values of y_{it} for which r_{it}=1.

Next, we follow Section 2.2.2 to re-organize this response indicator vector as well as the response vector itself. This allows for use of the estimating equations (2.25), (2.26) and (2.27) for the estimation of β, the autocorrelations and the variance of the regression estimates, respectively.

3.3.2 Simulation Results under MAR Models 1 and 2

For the estimation of the parameters, the data generated in the last sub-section along with covariates are now used in (2.25), (2.26) and (2.27) to estimate the regression parameter β, autocorrelations and the covariance matrix of the regression estimates, respectively.

Note that for the non-monotonic case, it was necessary to re-organize data as in Section 2.2.2. The whole estimation procedure was repeated for 1,000 simulated runs. The estimates of the parameters and the statistics (SM, SSE, SMSE and ESE) under different ρ values using D_1 and D_2 under **M1** and **M2** are reported in Tables A.4 and A.5 for the monotonic and non-monotonic cases respectively.

Note that under the monotonic case we have considered T=6 for **M1** and **M2**. This is done to compare the estimates obtained under MCAR models discussed in the last section. For the non-monotonic case, we have considered T=4 only. This is because the computation of the shifted vectors and matrices gets complicated for larger T . These results are comparable with MCAR cases with T=4 .

(a) Monotonic Case

It is clear from Table A.4 for the monotonic cases that, as expected, the estimates of the components of the β vector perform well for both models **M1** and **M2** under D_1 as compared to D_2. This is because D_2 contains time dependent covariates. For example, for true parameter values β_1=β_2=1, the estimates of the components of the β vector under **M1** are found to be 1.0371 and 1.0294 for D_1, with ρ=0.5, and 1.0451 and 1.0277 under **M2**, whereas, with the same ρ=0.5 for D_2, the estimates of the components of the β vector are found to be 0.9254 and 1.9805 under **M1**, and 0.9142 and 1.9901 under **M2**. We note that the β_2 estimate under D_2 also differs from that of D_1 considerably. This is due to the use of the time dependent covariate. Also as the ρ value increases, irrespective of the models (**M1** or **M2**), the estimates of the components of the β vector under D_1 appear to be unbiased, whereas the estimates under D_2 become slightly less biased.

Also, Table A.4 reveals that the correlation estimates approximately satisfy the AR(1) relationship. For example, under D_1, the correlation estimates are 0.79, 0.62, 0.48, 0.38 and 0.30, whereas the true correlation based on $\rho_\ell = \rho^\ell$ for ρ=0.8 are 0.8, 0.64, 0.51, 0.41 and 0.33 respectively.

We now provide a comparison of various estimates under **M1** and **M2** for D_1 only.

The SSE and ESE values under both **M1** and **M2** are close to each other. For example, under **M1** with ρ=0.5, the SSEs of the estimates of β_1 and β_2 are found to be 0.28 and 0.25 respectively, while their ESEs are found to be 0.27 and 0.24 respectively. Similarly, under **M2**, the SSEs of the estimates of β_1 and β_2 are found to be 0.29 and 0.26 respectively, while their ESEs are found to be 0.27 and 0.24 respectively. The SMSEs for β_1 and β_2 estimates increase as the value of ρ becomes larger. This is true under both models **M1** and **M2**. Also, the SMSEs under **M1** is smaller than that of **M2** as expected. For example with ρ=0.5 , the SMSEs for

β_1 and β_2 estimates are found to be 0.0806 and 0.0620 under **M1**, and 0.0876 and 0.0679 under **M2** respectively. Similarly, for ρ=0.9, the SMSEs of the estimates of β_1 and β_2 are found to be 0.1804 and 0.1400 under **M1**, and 0.1853 and 0.1457 under **M2**. This shows that the SMSEs are smaller under **M1** than **M2** in general. This also shows that the SMSEs increase for each β parameter as ρ value increases.

Note that when the MAR models (Table A.4) are compared with the MCAR models (Table A.2) under the monotonic case, the GQL approach performs better in β estimation under the MCAR model. This is because the values of the SMSEs are smaller under the MCAR case as compared to the MAR case. For example consider the MCAR model with $T = 6$, ρ=0.5, NMP=0.95 , the values of $\hat{\beta}_{1G}$ and $\hat{\beta}_{2G}$ were found to be 1.0207 and 1.0197 under D_1, while the SMSEs of the estimates of the β components were found to be 0.0500 and 0.0388 respectively. For similar parameter values under the same design, the estimates of β_1 and β_2 were found to be 1.0371 and 1.0294 respectively, while their SMSEs are 0.0806 and 0.0620 under the MAR model 1. Similarly, the SMSEs appear to be smaller under the MCAR models than under the MAR models for other values of ρ .

(b) Non-Monotonic Case

We now look at the results in Table A.5 which were computed under the non-monotonic pattern. As mentioned before, we consider T=4 only. Similar to the monotonic case, the estimation appears to work better under D_1 as compared to D_2. Here, for convenience, we discuss some of the estimation results under D_1 only.

It is clear from Table A.5 that the estimates of the components of the β vector perform well under models **M1** and **M2**. For example, with ρ=0.5, the estimates of the components of the β vector were found to be 1.0340 and 1.0108 under **M1** and 1.0353 and 1.0117 under **M2**, whereas the true values are 1.00 for both β parameters. Thus, the estimates are found to be unbiased.

Also Table A.5 reveals that the correlation estimates approximately satisfy the AR(1) relationship. For example, when the true value of ρ=0.8, the true three lag correlation estimates are 0.79, 0.65 and 0.54, whereas the actual lag correlations based on $\rho_\ell = \rho^\ell$ are 0.8, 0.64, 0.51 respectively.

The values of SSE and ESE under both **M1** and **M2** are approximately close to each other. For example, under **M1** with ρ=0.5, the SSEs are found to be 0.2694 and 0.2476 for the two β components, while the corresponding ESEs are 0.2731 and 0.2419. Similarly, under **M2**, the SSEs of the estimates of β_1 and β_2 are found to be 0.2661 and 0.2470, while their ESEs are 0.2731 and 0.2421 respectively. Thus the formulas for variance estimates work quite well. It was however observed that the SMSE under **M2** for the estimates of the components of the β vector appear to be smaller than that of under **M1**. For example, with ρ=0.8, the SMSEs of the estimates of the β components are found to be 0.1352 and 0.1033 under **M1**, and 0.1163 and 0.0925 under **M2**. But as the ρ value becomes larger, the SMSE under **M1** becomes smaller than that of **M2**. Note that when the MAR model (Table A.5) is compared with the MCAR model (Table A.3) under the non-monotonic case, the MCAR performs better in β estimation as expected. This is because SMSE values are smaller for the MCAR case as compared to the MAR case. For example, consider the non-monotonic MCAR model with $T = 4$, ρ=0.5 and NMP=0.95 under D_1, the values of $\hat{\beta}_{1G}$ and $\hat{\beta}_{2G}$ are 1.0201 and 1.0320 respectively, while their SMSEs are found to be 0.0637 and 0.0545 respectively. For similar parameter values under the same design, the estimates of the β components are found to be 1.0340 and 1.0108 respectively, and the corresponding SMSEs are found to be 0.0737 and 0.0614 under the MAR **M1** model. This shows that the SMSEs are smaller under the MCAR model as compared to the **M1** based MAR model.

(c) Overall Comparison

The GQL approach was applied to three different sets of data for the estimation of regression parameters. To be specific, we have generated longitudinal data under a complete model as well as under two longitudinal missing models, namely, MCAR and MAR models, and the GQL approach was subsequently applied to all three sets of data to estimate the same β parameter.

A comparison for the performance of the GQL approach in estimating β based on complete longitudinal data and the data generated under MCAR and MAR models indicates that the GQL approach performs better when it is applied to the complete data as expected. For example, under the complete model and under design D_1, the SM values for $\hat{\beta}_{1G}$ and $\hat{\beta}_{2G}$ are 1.0221 and 1.0207 with their corresponding SSEs as 0.2113 and 0.1941 for the case T=6 and ρ=0.5 respectively, the ESEs are found to be 0.2114 and 0.1863, and the corresponding SMSEs are found to be 0.0451 and 0.0384 . For the monotonic MCAR case with NMP=0.95, using the same design and ρ as in the complete case, the SM values for $\hat{\beta}_{1G}$ and $\hat{\beta}_{2G}$ are found to be 1.0270 and 1.0197 respectively, their SSEs are 0.2225 and 0.1960, the ESEs are found to be 0.2133 and 0.1905, while the corresponding SMSEs are found to be 0.0500 and 0.0388 respectively. Under the monotonic MAR model 1 (**M1**), the SM values are 1.0371 and 1.0294 for $\hat{\beta}_{1G}$ and $\hat{\beta}_{2G}$ respectively, their SSEs are found to be 0.2815 and 0.2473, the ESEs are found to be 0.2702 and 0.2396, while the corresponding SMSEs are found to be 0.0806 and 0.0620 respectively. Next, for the monotonic MAR model 2 (**M2**) with $\gamma_1 = 0.3$ and $\gamma_2 = 0.7$, the SM values are 1.0451 and 1.0277 for $\hat{\beta}_{1G}$ and $\hat{\beta}_{2G}$ respectively, their SSEs are found to be 0.2925 and 0.2591, the ESE are found to be 0.2709 and 0.2402, while the corresponding SMSEs are found to be 0.0876 and 0.0679 respectively.

In summary, for the parameters described above, the complete data based GQL approach is found to be 1.1 times more efficient than the MCAR model based estimation in estimating the same parameter. Similarly, the complete data based GQL approach is found to be 1.4 times more efficient than the MAR (**M1**) based estimation, and 1.5 times more efficient than the MAR (**M2**) based estimation in estimating β_2. This provides a clear idea about the loss of efficiency of the GQL approach for the analysis of the missing data as compared to the complete data. More specifically, the GQL approach does not appear to lose any efficiency if the data is MCAR with high non-missing probability (NMP). The GQL approach however can be inefficient if the missing data follow MAR models, especially the MAR **M2** models.

Chapter 4

Analysis of the SLID (Survey of Labour and Income Dynamics) Data in the Presence of Missing Responses

4.1 Introduction to the SLID Data

In this section, we revisit the SLID data that was analyzed earlier by Sutradhar and Kovacevic (2003). To be specific, we consider a subset of the SLID data that was collected by Statistics Canada for the period from 1995-1998. Sutradhar and Kovacevic (2003) considered a longitudinal binary response data set for these six years. The binary variable was 'unemployed all year', derived from a variable 'labour force status for the year', assigns value 1 to those who were unemployed for the entire year, and 0 to those who were employed for the full year or a part of year employed and part unemployed. A missing response for a person who contributed a response for at least one year was considered as a missing value of the response variable, although a person could have left the labour force. Sutradhar and Kovacevic (2003) have identified 18,077 respondents in the domain of interest. Among them 15,731 individuals were found to have complete data for all six years, and the remaining 2,346 individuals had missing responses in a monotonic pattern. These authors have however analyzed this longitudinal data set under the assumption that the missingness occurred completely at random. In this section, unlike these authors, we assume that the missing responses occurred at random. Thus we analyze the same data set as in Sutradhar and Kovacevic (2003) but under a MAR mechanism.

For convenience, we describe the data set before we undertake the confirmatory analysis. As far as the pattern of the missing data is concerned, we consider a monotonic mechanism only since the study gives rise to missing data in this fashion.

The SLID response data in monotonic missing form is reported in Table 4.1. More specifically, the first block (year 1993) was recorded for all individuals at the start of the study, and hence it is completely observed. The second block (year 1994) consists of responses from 17576 individuals with 97.23% observed in the follow-up study, the 3rd block (1995) consist of response from 17000 individuals with 94.04% observed. Block 1 contains more observations than block 2. Similarly, block 2 contains more observations that block 3, and so on. Thus, the blocks form a monotone pattern of missing data.

As Table 4.1 shows, the number of individuals with 1, 2, 3, 4, and 5 missing values were found to be 413, 460, 396, 576, and 501 respectively. The number of unemployed individuals appear to increase to 408 in 1994 from 359 in 1993. The unemployed number however has decreased since 1995. The purpose of this study is to examine the effects of the associated characteristics or covariates on employment status by taking the longitudinal correlation of the response as well as the missing pattern of the response into account. Some common characteristics that may be related to the longitudinal all-year unemployment data are: gender, age, geographical location, education level, and marital status of the individual. Note that the binary responses collected over six years are longitudinally correlated. Also, some of them are missing. To address the purpose of this study, we find the effects of the 5 main

Table 4.1: Sample counts of 'unemployed' and distribution of missing values over time

					Year		
Response Status	Unemployment Status & Missing frequency	1993	1994	1995	1996	1997	1998
Complete	Employed(=0)	17718	17168	16623	16235	15824	15455
	Unemployed(=1)	359	408	377	369	320	276
Percent of Complete		100	97.23	94.04	91.85	89.31	87.02
Missing	Once		501	576	396	460	413
	Twice			501	576	396	460
	Three times				501	576	396
	Four times					501	576
	Five times						501
	Total Missing	0	501	1077	1473	1933	2346
Percent of Missing		0	2.77	5.96	8.15	10.69	12.98
Total Individuals		18,077	18,077	18,077	18,077	18,077	18,077

covariates (characteristics) on all-year unemployment after taking the longitudinal and missingness nature of the responses into account.

To shed some light on the nature of the longitudinal relationship between the binary responses 'unemployed all year' and the 5 covariates, we construct appropriate 3-way tables for the 5 covariates and the binary response variable 'unemployed all-year' for the duration from 1993 to 1998. At each level of the selected covariates, we also exhibit the number of missing values over time, that is, the number of individuals having no response. These results are reported in Tables 4.2 to 4.6, for the age, gender, region of residence, education level, and marital status respectively.

Based on the complete data, it is clear from Table 4.2 that there are more unemployed individuals in the age group of 25 to 55 which is obvious as this group has the largest range. The proportions of unemployed individuals are however also larger for this group followed by the 16 to 25 age group. The older age group 55 to 65 has the smallest proportions of unemployment from 1994 to 1998. The proportion unemployed appears to decrease over time in all three groups since 1994. With regard to the frequency of missing responses, the youngest age group has the largest nonresponse rate among the 3 age groups.

Table 4.3 shows that the proportion of unemployed females is generally more than that of males. Specifically, over the years 1994-98, there were more unemployed females than males. As far as the missing values are concerned, the number of nonrespondent male is seen to be larger as compared to the females. This is true for all 5 years from 1994 to 1998.

Table 4.4 shows that the proportion unemployed is the highest in Atlantic region followed by Quebec, Ontario, BC & Alberta, and Prairies. Note that the proportion unemployed in BC & Alberta is only slightly higher than Prairies. Similarly the proportions unemployed in the Atlantic region is slightly higher than Quebec except

Table 4.2: Sample counts cross-classified according to 'unemployed' and 'age' group in 1993

				Year			
Age group	Unemployment Status	1993	1994	1995	1996	1997	1998
$16 \leq$ Age in 1993< 25	Employed(=0)	2978	2816	2667	2543	2412	2316
	Unemployed(=1)	51	69	68	62	46	37
	Missing	0	144	294	424	571	676
$25 \leq$ Age in 1993< 55	Employed(=0)	12385	12037	11690	11449	11199	10960
	Unemployed(=1)	250	290	271	273	247	216
	Missing	0	308	674	913	1189	1459
$55 \leq$ Age in 1993< 65	Employed(=0)	2355	2315	2266	2243	2213	2179
	Unemployed(=1)	58	49	38	34	27	23
	Missing	0	49	109	136	173	211

Table 4.3: Sample counts cross-classified according to 'unemployment status' and 'sex'

					Year		
Sex	Unemployment Status	1993	1994	1995	1996	1997	1998
Male	Employed(=0)	8547	8286	7996	7769	7559	7373
	Unemployed(=1)	175	177	168	175	151	123
	Missing	0	259	558	778	1012	1226
Female	Employed(=0)	9171	8882	8627	8466	8265	8082
	Unemployed(=1)	184	231	209	194	169	153
	Missing	0	242	519	695	921	1120

Table 4.4: Sample counts cross-classified by 'Region of residence' and 'Unemployed'

					Year		
Region of residence	Unemployment status	1993	1994	1995	1996	1997	1998
Atlantic	Employed(=0)	3878	3752	3652	3548	3445	3366
	Unemployed(=1)	113	124	117	131	109	93
	Missing	0	80	167	236	330	385
Quebec	Employed(=0)	3596	3493	3407	3367	3309	3233
	Unemployed(=1)	94	119	121	107	88	79
	Missing	0	79	159	209	284	358
Ontario	Employed(=0)	4444	4284	4180	4069	3941	3862
	Unemployed(=1)	91	87	73	81	76	59
	Missing	0	181	309	429	568	703
Prairies	Employed(=0)	4260	4122	3893	3785	3700	3603
	Unemployed(=1)	44	58	50	36	33	27
	Missing	0	107	343	453	574	690
BC & Alberta	Employed(=0)	1540	1517	1491	1466	1429	1391
	Unemployed(=1)	17	20	16	14	14	18
	Missing	0	54	99	146	177	210

for 1994 and 1995, Ontario appears to have middle place in the country with regard to the unemployment status of the individuals. With regard to the proportion of nonresponse, the province of Ontario appears to have the largest non-response rate of 2190 (29.9%) followed by Prairies with 2167(29.6%).

Table 4.5 helps us to understand the effect of education on unemployment over the years. It is clear from the above table that the 'high education' group has the smallest unemployment rate followed by the 'medium education' group, as expected. The

Table 4.5: Sample counts cross-classified according to 'Education level' and 'Unemployed'

					Year		
Education level	Unemployment status	1993	1994	1995	1996	1997	1998
Low education	Employed(=0)	3708	3320	3121	3002	2896	2821
	Unemployed(=1)	140	133	132	121	115	96
	Missing	0	102	203	278	341	405
Medium education	Employed(=0)	11731	11521	11136	10836	10488	10151
	Unemployed(=1)	203	251	229	231	185	168
	Missing	0	344	750	1018	1349	1627
High education	Employed(=0)	2279	2327	2366	2397	2440	2483
	Unemployment(=1)	16	24	16	17	20	12
	Missing	0	55	123	176	241	311

unemployment proportions are quite high over the years in the 'low education' group. Once again, similar to other covariates, the unemployment rates corresponding to this 'education level' also appear to increase in 1994 from 1993 but start decreasing slowly beginning from 1995. The 'high education' group has the smallest nonresponse rate which is also expected.

Table 4.6 shows that the proportion of unemployed individuals is smaller over the years in the 'married/common law' group, followed by 'widowed' , 'single' and 'separated/divorce' groups. More specifically, the proportions are closer between the 'married/common law' and 'widowed' groups, and also between the 'single' and the 'separated/divorced' groups. But when the 'married/common law' or 'widowed' group is compared with 'single' or 'separated/divorced' group, their proportions appear to be quite different. Both of the 'separated/divorced' and 'single' groups also appear to have higher nonresponse rates all throughout the years.

Table 4.6: Sample counts cross-classification by 'Marital status' and 'Unemployed'

					Year		
Marital status	Unemployment status	1993	1994	1995	1996	1997	1998
Married/common law	Employed(=0)	12020	11800	11607	11566	11430	11305
	Unemployed(=1)	214	246	198	199	176	143
	Missing	0	266	593	810	1069	1342
Separated/divorced	Employed(=0)	1188	1281	1321	1384	1393	1426
	Unemployed(=1)	44	41	65	56	57	48
	Missing	0	38	129	187	263	342
Widowed	Employed(=0)	330	365	393	410	437	465
	Unemployed(=1)	7	6	8	10	5	6
	Missing	0	4	21	33	42	54
Single	Employed(=0)	4180	3722	3302	2875	2564	2259
	Unemployed(=1)	94	115	106	104	82	79
	Missing	0	193	334	443	559	608

4.2 Notation for the SLID Data Analysis

In this section, we denote the response and the covariates of the SLID data by using our notation provided in Chapter 2, for example. To be specific, we denote the binary response variable 'unemployed all year' by y_{it} for $i = 1, \ldots, 18077$ and $t = 1, \ldots, T_i$, where T_i denotes the number of response available for the ith individual with its range $T_i \leq T$=6.

As far as the covariates are concerned, as they are independent of time, we rename the 5 covariates discussed in section 4.1 as follows: First, gender is represented by x_1 which is 0 for female and 1 for male. The second covariate 'age' is represented by x_2 in general. To be specific, we consider 3 age groups based on their ages at 1993: group 1 consists of individuals between 16 and 24 inclusive , group 2 consists of individuals between 25 and 54, and group 3 from 55 to 65. Now by considering the younger age group 1 as the referenced group, we represent the above 3 groups by x_{21} and x_{22}, so that x_{21}=0 , x_{22}=0 stands for the individual of the group 1, x_{21}=1, x_{22}=0 represent the individual of the group 2, and x_{21}=0, x_{22}=1 would identify the individual belonging to group 3.

The third covariate 'education level' is represented by x_3. To be specific, we consider x_{31} and x_{32} to represent 3 levels (low, medium and high) of education, lower level being the reference level, say. Thus, x_{31}=0 and x_{32}=1 will represent an individual with high education level.

The fourth covariate 'marital status' is denoted by x_4. As the marital status can be married & common law spouse, separated & divorce, widow, or single (never married), we use 3 covariates x_{41}, x_{42}, and x_{43} respectively to represent them , married and common law spouse group being the reference group. Finally, we consider x_5 to represent geographical location, where x_{51}, x_{52}, x_{53}, and x_{54} are covariates used to identify an individual from any of the Atlantic, Quebec, Ontario, Praries, or British columbia & Alberta regions. Here we consider the Atlantic region as the reference region with all 4 covariates as 0; x_{51}=1, $x_{52} = x_{53} = x_{54}$=0 will represent the individual from Quebec, and so on.

Note that altogether there are 12 covariates. In Sections 4.3 and 4.4, we compute the effects of these covariates on the binary all-year unemployment variable after taking the longitudinal correlations of the data as well as the missingness pattern into account.

Although interaction may be possible within the covariates, but in this practicum, we only consider the simple linear case.

4.3 Incomplete SLID Data Analysis When Some Longitudinal Responses are monotonically MAR Following M1 or M2

Sutradhar and Kovacevic (2003) analyzed the same data under the complete and MCAR cases, therefore, we do not compute the effects of the covariates under these complete and MCAR models as the results are available in their paper.

In this section, we compute the effects of all 12 covariates under the assumption that missing indicators follow either **M1** or **M2** models. Recall that under the **M1** model , we consider logit $\{Pr(r_{it} = 1)\} = y_{i1}$, and similarly, under the **M2** model, we consider logit $\{Pr(r_{it} = 1)\} = \gamma_1 y_{i1} + \gamma_2 y_{i2}$. These non-response probability structures then help us to write the formula for δ_{it} given by

$$\delta_{it} = r_{it}/Pr\{(\prod_{j=1}^{t} r_{ij}) = 1 | H_{i,t-1}, \gamma\}$$

as in (2.21). Next, we re-express the mean of the binary response as

$$E(y_{it}) = \mu_{it} = a'(\theta_{it}) = exp(\theta_{it})/[1 + exp(\theta_{it})]$$

where $\theta_{it} = x'_{it}\beta$, x'_{it} ($=x'_{i.}$ for all t) being the 1×12 vector representing all 12 covariates generated from the 5 original covariates as discussed.

The above expression for δ_{it} and μ_{it} are then used in (2.26) to compute the longitudinal correlations ρ_ℓ for $\ell = 1, \ldots, 5$. Note that y_{it} values involved in z_{it} for (2.26) are observed responses for all $i = 1, \ldots, 18077$, and $t = 1, \ldots, T_i \leq 6$.

The correlation estimates along with the longitudinal weights δ_{it} are then used in (2.25) to compute the regression estimates under the **M1** or **M2** models. Note that the estimates of β and $\rho_\ell(\ell = 1, \ldots, 5)$ are obtained iteratively. The estimates of these parameters are reported in Table 4.7 under MAR **M1** model and in Table 4.8 under MAR **M2** model. Note that to compute the non-response probability under the MAR **M2** model, we have considered γ_1=0.3 and γ_2=0.7, so that more weights are given on the recent observation between y_{i1} and y_{i2}. This selection is however, subjective, which could be avoided by estimating these parameters from the data. This is however beyond the scope of the present practicum.

In order to be able to construct the confidence intervals for the estimates of the regression effects, we have also computed their standard errors by using the formula (2.27) for the covariance matrix of $\hat{\beta}_{GQL}$. These results are also reported in Table 4.7 under the **M1** model and in Table 4.8 under the **M2** model.

For the analysis of the SLID data under **M1** model, we deal with all 18,077 individuals, as the response y_{i1} is available for them. From the result of Table 4.7, the longitudinal correlations appear to be moderate and decay as the time lag increases. With regard to the interpretation of the regression effects, the negative value - 0.1324 for the gender effects indicates that a male has lower probability of an all-year unemployment as compare to the female. The negative values -1.1492 and -1.8152 of β_2 and β_3 indicate that the younger group has higher probability of an all-year unemployment and the probability decreases for older age groups.

As far as the effect of geographic location on the all-year unemployment is concerned, it appears that Quebec had the smallest probability of an all-year unemployment during 1993 to 1998 followed by BC & Alberta, Praries, Ontario and Atlantic provinces. This follows from the fact that the regression estimates for Quebec, Ontario, BC & Alberta, and Praries are found to be -0.5897 , -0.0353 , -0.4384 and

-0.0864 respectively.

The larger negative value -1.0749 for β_5 as compared to β_4= -0.7140 indicates that as the education level gets higher, the probability of an all-year unemployment gets smaller. Finally, with regard to marital status, the positive value 0.1405 for β_6 means that the separated and divorced individuals have higher probability of all-year unemployment as compared to the married and common law spouse group. Similarly, a widow had less probability of all-year unemployment as compared to a single individual.

We can also interpret the result by using the odds ratios. For example, the odds ratio for Quebec is found to be 0.55, this implies that the odds of observing an unemployed individual from Quebec is less likely as compared to the odds of observing an unemployed individual in the Atlantic region, given that all other covariates remain fixed.

For the analysis of the SLID data under **M2** model, we assume that the first two responses of an individual must be available in order to include the individual in the study. This is because under **M2** model the response indicator variable $r_{it}(t = 3, \ldots, 6)$ is dependent on y_{i1} and y_{i2} for the ith individual. The regression estimates along with their standard errors, and also the values of the longitudinal correlations, are reported in Table 4.8. The longitudinal correlation estimates under both **M1** and **M2** models appear to be quite similar. As for the estimation of the main parameters, the GQL approach in general produces similar estimates under both **M1** and **M2** models, except for example, x_{42}(marital status 3 vs 1) is found to be -0.22 under **M2** model but 0.08 under **M2** model. This means under **M1** model, a widow has better chance of being employed whereas **M2** model increases the probability for unemployment. The standard errors of the regression estimates under **M1** model were however found to be smaller than that of **M2** model. This pattern is also supported by our simulation studies, where it was found that **M1** model produces estimates with smaller standard error. As the regression estimates are generally similar and standard errors under **M1** model are smaller, we recommend the use of **M1** models between **M1** and **M2** models, for the analysis of the SLID data.

Note however that when the regression estimates along with their standard errors provided in Table 4.7 and 4.8 are compared with corresponding values under the MCAR and complete models as given in Sutradhar and Kovacevic (2003), the latter models produce estimates with smaller standard errors, which is expected.

Table 4.7: Estimates of regression and their Estimated Standard Errors, as well as estimates of autocorrelations for the SLID data with MAR **M1** type nonresponse

Parameters	Estimate	Standard Error
Male vs Female(x_1)	-0.1324	0.0387
Age group 2 vs 1(x_{21})	-1.1492	0.0412
Age group 3 vs 1(x_{22})	-1.8152	0.0775
Education Med. vs low(x_{31})	-0.7140	0.0387
Education high vs Low(x_{32})	-1.0749	0.0768
Marital Status 2 vs 1(x_{41})	0.1405	0.0678
Marital Status 3 vs 1(x_{42})	-0.2241	0.1673
Marital status 4 vs 1(x_{43})	0.1101	0.0447
Quebec vs Atlantic (x_{51})	-0.5897	0.06
Ontario vs Atlantic (x_{52})	-0.0353	0.0529
Praries vs Atlantic (x_{53})	-0.0864	0.0529
BC & Alberta vs Atlantic (x_{54})	-0.4384	0.0831
ρ_1	0.6802	
ρ_2	0.5599	
ρ_3	0.4058	
ρ_4	0.2334	
ρ_5	0.0073	

Table 4.8: Estimates of regression and their Estimated Standard Errors, as well as estimates of autocorrelations for the SLID data with MAR **M2** type nonresponse with γ_1=0.3 and γ_2=0.7

Parameters	Estimate	Standard Error
Male vs Female(x_1)	-0.0407	0.0424
Age group 2 vs 1(x_{21})	-1.2377	0.0447
Age group 3 vs 1(x_{22})	-1.8240	0.0794
Education Med. vs low(x_{31})	-0.9224	0.0412
Education high vs Low(x_{32})	-1.3900	0.0866
Marital Status 2 vs 1(x_{41})	0.1709	0.0721
Marital Status 3 vs 1(x_{42})	0.0839	0.1559
Marital status 4 vs 1(x_{43})	-0.2842	0.0510
Quebec vs Atlantic (x_{51})	-0.6223	0.0640
Ontario vs Atlantic (x_{52})	-0.0969	0.0557
Praries vs Atlantic (x_{53})	-0.2367	0.0574
BC & Alberta vs Atlantic (x_{54})	-0.6255	0.0949
ρ_1	0.6655	
ρ_2	0.4915	
ρ_3	0.2362	
ρ_4	0.028	
ρ_5	0.072	

Chapter 5

Conclusion

RRZ (1995) and Paik (1997) have extended the GEE approach of Liang and Zeger (1986) to accommodate the longitudinal data analysis with outcomes subject to non-response. As these approaches use the so-called 'working' correlations chosen by the investigator, they may not always yield efficient estimates for regression parameters. This raised an issue of using a robust correlation structure for the longitudinal data in order to construct estimating equations for the purpose of obtaining consistent and efficient regression estimates. Following Sutradhar and Das (1999)[see also Jowaheer and Sutradhar (2002)], Sutradhar and Kovacevic (2003) have developed a general correlation structure based GQL (generalized quasi-likelihood) approach to analyse the longitudinal missing data subject to MCAR and weighted GQL (WGQL) approach to analyze longitudinal missing data subject to MAR. These authors then have applied their estimation methodology to analyze SLID data under the assumption that the missing longitudinal responses are subject to MCAR only.

In this practicum, we have examined the performance of the GQL and WGQL approaches of Sutradhar and Kovacevic (2003) through a simulation study. More specifically, to begin with, we have found that the GQL approach is more efficient than the 'working' independence approach in estimating regression coefficients under a complete model. This was examined by considering an AR(1) correlation model for the complete longitudinal data. Next it was found that the GQL approach performs well in estimating the regression effects under the MCAR model provided the non-response probabilities are not too low. Similarly, the WGQL approach was found to work well under a less restricted MAR (**M1**) model.

When performance of the WGQL approach for the MAR models was compared with the GQL approach for the MCAR model, the latter was found to be more efficient (in the sense of lower mean squared error), as expected. The MAR model based estimation methodology was also applied to reanalyze the SLID data that was earlier analyzed by Sutradhar and Kovacevic (2003) under the MCAR model. Remark that as the true longitudinal response of the SLID data is not known, it would be more appealing to develop some statistical tests to detect the non-response mechanism in order to provide an improved estimation methodology. This however appears to be a difficult problem and beyond the scope of the present practicum.

Appendix A

Simulation Result

Table A.1: **Non-Missing Case**: Simulated means (SM), simulated standard errors (SSE), simulated mean square error (SMSE), and estimated standard error (ESE) of the regression estimators based on GQL and GEE(I) approaches; SM and SSE of moment estimates for longitudinal correlation parameter under binary AR(1) process with T=6,10 and 15, K=100, $\beta_1 = \beta_2 = 1$; based on 1000 simulations.

$T = 6$ $Design : D_1$

ρ	Statistic	$\hat{\beta_{1I}}$	$\hat{\beta_{2I}}$	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
0.5	SM	1.0306	1.0286	1.0221	1.0207	0.4924	0.2414	0.1207	0.0602	0.0298
	SSE	0.2684	0.2443	0.2113	0.1941	0.0504	0.0670	0.0725	0.0836	0.1047
	ESE	0.1249	0.1105	0.2114	0.1863					
	SMSE	0.0730	0.0605	0.0451	0.0384					
0.8	SM	1.0263	1.0320	1.0284	1.0390	0.7951	0.6314	0.5010	0.3959	0.3155
	SSE	0.3161	0.2862	0.2964	0.2676	0.0404	0.0679	0.0862	0.1011	0.1159
	ESE	0.1655	0.1473	0.2784	0.2478					
	SMSE	0.1006	0.0829	0.0887	0.0731					
0.9	SM	1.0358	1.0440	1.0487	1.0605	0.8964	0.8034	0.7195	0.6449	0.5772
	SSE	0.3378	0.3139	0.3375	0.3092	0.0284	0.0522	0.0729	0.09085	0.1069
	ESE	0.1841	0.1640	0.3184	0.2864					
	SMSE	0.1154	0.1005	0.1163	0.0993					

$T = 6$ $Design : D_2$

ρ	Statistic	$\hat{\beta_{1I}}$	$\hat{\beta_{2I}}$	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
0.5	SM	0.9995	1.8804	0.9583	1.2908	0.4993	0.2455	0.1113	0.0353	-0.0128
	SSE	0.2578	0.3273	0.2042	0.2496	0.0479	0.0639	0.0755	0.0867	0.1125
	ESE	0.1204	0.0742	0.2030	0.2173					
	SMSE	0.0665	0.8822	0.0434	0.1469					
0.8	SM	1.0076	1.8802	0.9535	0.9233	0.7984	0.6377	0.5077	0.4029	0.3195
	SSE	0.3017	0.3688	0.2660	0.2898	0.0369	0.0643	0.0870	0.1043	0.1225
	ESE	0.1597	0.0980	0.2553	0.2170					
	SMSE	0.0911	0.9108	0.0729	0.0899					
0.9	SM	1.0076	1.8802	0.9257	0.5684	0.9036	0.8179	0.7399	0.6699	0.6068
	SSE	0.3017	0.3688	0.2913	0.2512	0.0251	0.0479	0.06944	0.08782	0.1047
	ESE	0.1597	0.0980	0.2775	0.1780					
	SMSE	0.0911	0.9108	0.0904	0.2494					

$T = 10 \quad Design : D_1$

ρ	Statistic	$\hat{\beta_{1I}}$	$\hat{\beta_{2I}}$	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
0.5	SM	1.0073	1.0179	1.0041	1.0109	0.4965	0.2458	0.1221	0.0583	0.0264
	SSE	0.2461	0.2304	0.1675	0.1562	0.0403	0.0496	0.0523	0.0530	0.0569
	ESE	0.0768	0.0678	0.1709	0.1517					
	SMSE	0.0606	0.0534	0.0281	0.0245					
0.8	SM	1.0148	1.0221	1.0149	1.0185	0.7965	0.6338	0.5044	0.3993	0.3156
	SSE	0.2838	0.2629	0.2467	0.2288	0.0317	0.0524	0.0666	0.0768	0.0837
	ESE	0.1122	0.1005	0.2421	0.2156					
	SMSE	0.0808	0.0696	0.0611	0.0527					
0.9	SM	1.0234	1.0327	1.0319	1.0443	0.8964	0.8026	0.7187	0.6436	0.5772
	SSE	0.3104	0.2942	0.2955	0.2825	0.0238	0.0435	0.0584	0.0714	0.0823
	ESE	0.1319	0.1175	0.2913	0.2617					
	SMSE	0.0969	0.0876	0.0883	0.0818					

$T = 10 \quad Design : D_2$

ρ	Statistic	$\hat{\beta_{1I}}$	$\hat{\beta_{2I}}$	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
0.5	SM	0.8661	1.7719	0.9443	1.3745	0.5070	0.2588	0.1316	0.0626	0.0218
	SSE	0.2183	0.1875	0.1647	0.2079	0.0363	0.0468	0.0493	0.0533	0.0583
	ESE	0.0707	0.0346	0.1667	0.1910					
	SMSE	0.0656	0.6310	0.0302	0.1835					
0.8	SM	0.8703	1.7712	0.9259	1.0776	0.8007	0.6409	0.5131	0.4088	0.3242
	SSE	0.2498	0.1998	0.2380	0.2808	0.0268	0.0460	0.0607	0.0730	0.0826
	ESE	0.1000	0.0480	0.2274	0.2223					
	SMSE	0.0792	0.6347	0.0621	0.0849					
0.9	SM	0.8680	1.7792	0.9124	0.7608	0.9023	0.8145	0.7357	0.6654	0.6021
	SSE	0.2712	0.2137	0.2689	0.2709	0.0194	0.0364	0.0561	0.0648	0.0765
	ESE	0.1153	0.0557	0.2592	0.1987					
	SMSE	0.0910	0.6528	0.0800	0.1306					

$T = 15 \quad Design : D_1$

ρ	Statistic	$\hat{\beta_{1I}}$	$\hat{\beta_{2I}}$	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
0.5	SM	1.1502	1.003	0.9989	1.0146	0.4967	0.2455	0.1192	0.0557	0.0256
	SSE	0.6849	0.4680	0.1476	0.1276	0.0316	0.0398	0.0410	0.0406	0.0415
	ESE	0.0224	0.0283	0.1435	0.1265					
	SMSE	0.4916	0.2190	0.0218	0.0165					
0.8	SM	1.1132	0.9988	1.0035	1.0166	0.7972	0.6352	0.5052	0.4015	0.3186
	SSE	0.6734	0.2005	0.2320	0.2005	0.0237	0.0393	0.0498	0.0568	0.0616
	ESE	0.0469	0.0490	0.2135	0.1884					
	SMSE	0.4663	0.0402	0.0538	0.0405					
0.9	SM	1.1049	1.0055	1.0297	1.0418	0.8967	0.8037	0.7200	0.6447	0.5772
	SSE	0.4790	0.3583	0.2773	0.2483	0.0190	0.0340	0.0468	0.0581	0.0674
	ESE	0.0735	0.0632	0.2701	0.2487					
	SMSE	0.2404	0.1284	0.0778	0.0634					

$T = 15 \quad Design : D_2$

ρ	Statistic	$\hat{\beta_{1I}}$	$\hat{\beta_{2I}}$	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
0.5	SM	1.0483	0.7308	0.9387	1.4339	0.5128	0.2670	0.1408	0.0751	0.0392
	SSE	0.2432	0.5623	0.1352	0.1798	0.0287	0.0359	0.0384	0.0399	0.0403
	ESE	0.0436	0.0265	0.1418	0.1685					
	SMSE	0.0615	0.3886	0.0220	0.2206					
0.8	SM	1.0439	0.7275	0.9250	1.2068	0.8016	0.6425	0.5147	0.4119	0.3296
	SSE	0.2421	0.5004	0.2039	0.2670	0.0215	0.0367	0.0476	0.0562	0.0631
	ESE	0.0548	0.0316	0.2035	0.2173					
	SMSE	0.0605	0.3247	0.0472	0.1141					
0.9	SM	1.0443	0.7502	0.9286	0.9271	0.9008	0.8114	0.7308	0.6585	0.5934
	SSE	0.2379	0.4339	0.2500	0.2842	0.0155	0.0285	0.0405	0.0517	0.0619
	ESE	0.0663	0.0346	0.2415	0.2126					
	SMSE	0.0586	0.2507	0.0676	0.0861					

Table A.2: **Monotonic MCAR Case**: Simulated means (SM), simulated standard errors (SSE), simulated mean square error (SMSE), and estimated standard error (ESE) of the regression estimators based on GQL approach; SM and SSE of moment estimates for longitudinal correlation parameter under binary AR(1) process for the case with T= 4 ; and non-missing probabilities (NMP) 0.80, 0.90 and 0.95 for T=6,10 and 15 ; K=100, $\beta_1 = \beta_2 = 1$; based on 1000 simulations.

$T = 4$ $NMP = 0.90$

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$
D_1	0.5	SM	1.0284	1.0278	0.4914	0.2361	0.1097
		SSE	0.2575	0.2236	0.0689	0.0928	0.1186
		ESE	0.2484	0.2241			
		SMSE	0.0671	0.0508			
	0.8	SM	1.0500	1.0449	0.7917	0.6235	0.4908
		SSE	0.3168	0.2923	0.0555	0.0957	0.1248
		ESE	0.3084	0.2782			
		SMSE	0.1029	0.0875			
	0.9	SM	1.0526	1.0467	0.8930	0.7948	0.7131
		SSE	0.3402	0.3126	0.0417	0.0822	0.1096
		ESE	0.3370	0.3018			
		SMSE	0.1185	0.0999			

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
D_2	0.5	SM	0.9804	1.2108	0.4921	0.2316	0.0854
		SSE	0.2474	0.2804	0.0661	0.0959	0.1274
		ESE	0.2392	0.2415			
		SMSE	0.0616	0.1231			
	0.8	SM	0.9380	0.7435	0.8012	0.6420	0.5116
		SSE	0.2899	0.3383	0.0506	0.0926	0.1288
		ESE	0.2832	0.2349			
		SMSE	0.0879	0.1802			
	0.9	SM	0.9410	0.5824	0.8922	0.7928	0.7051
		SSE	0.3074	0.2345	0.0316	0.0629	0.0906
		ESE	0.2982	0.2152			
		SMSE	0.0980	0.2294			

$T = 4 \quad NMP = 0.95$

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$
D_1	0.5	SM	1.0253	1.0346	0.4900	0.2339	0.1053
		SSE	0.2423	0.2279	0.0662	0.0871	0.1116
		ESE	0.2458	0.2216			
		SMSE	0.0593	0.0531			
	0.8	SM	1.0500	1.0522	0.7926	0.6252	0.4925
		SSE	0.3053	0.2891	0.0520	0.0860	0.1182
		ESE	0.3058	0.2760			
		SMSE	0.0957	0.0863			
	0.9	SM	1.0616	1.0697	0.8945	0.7979	0.7132
		SSE	0.3260	0.3197	0.0385	0.0691	0.0992
		ESE	0.3324	0.3023			
		SMSE	0.1100	0.1071			

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$
D_2	0.5	SM	0.9784	1.2105	0.4918	0.2323	0.0841
		SSE	0.2505	0.2686	0.0647	0.0915	0.1212
		ESE	0.2379	0.2358			
		SMSE	0.0632	0.1165			
	0.8	SM	0.9407	0.7702	0.8008	0.6408	0.5095
		SSE	0.2949	0.3062	0.0483	0.0872	0.1220
		ESE	0.2827	0.2263			
		SMSE	0.0904	0.1466			
	0.9	SM	0.9347	0.5585	0.8964	0.8026	0.7210
		SSE	0.2895	0.2075	0.0309	0.0609	0.0854
		ESE	0.2970	0.1980			
		SMSE	0.0881	0.2380			

T =6 NMP=0.80

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_1	0.5	SM	1.0162	1.0230	0.4933	0.2410	0.1194	0.0544	0.0222
		SSE	0.2491	0.2174	0.0615	0.0823	0.0947	0.1116	0.1448
		ESE	0.2345	0.2163					
		SMSE	0.0623	0.0478					
	0.8	SM	1.0485	1.0475	0.7894	0.6215	0.4915	0.3884	0.3096
		SSE	0.3096	0.2842	0.0540	0.0904	0.1153	0.1379	0.1659
		ESE	0.3065	0.2791					
		SMSE	0.0982	0.0830					
	0.9	SM	1.0928	1.0776	0.8899	0.7886	0.7022	0.6260	0.5637
		SSE	0.3844	0.3451	0.04471	0.0870	0.1168	0.1467	0.1731
		ESE	0.3770	0.3344					
		SMSE	0.1564	0.1251					

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_2	0.5	SM	0.9516	1.4206	0.5033	0.2485	0.1098	0.0240	-0.0390
		SSE	0.2333	0.3531	0.0603	0.0845	0.1047	0.1310	0.1767
		ESE	0.2254	0.2804					
		SMSE	0.0568	0.3016					
	0.8	SM	0.9400	0.9707	0.7932	0.6223	0.4822	0.3631	0.2615
		SSE	0.2781	0.4273	0.0449	0.0803	0.1111	0.1385	0.1762
		ESE	0.2773	0.3234					
		SMSE	0.0809	0.1834					
	0.9	SM	0.9564	0.7218	0.8842	0.7695	0.6645	0.5617	0.4783
		SSE	0.3116	0.3393	0.0298	0.0640	0.0951	0.1192	0.1423
		ESE	0.2994	0.3597					
		SMSE	0.0990	0.1925					

T =6 NMP=0.90

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_1	0.5	SM	1.0187	1.0252	0.4934	0.2403	0.1156	0.0543	0.0245
		SSE	0.2287	0.1986	0.0578	0.0748	0.0841	0.0941	0.1178
		ESE	0.2205	0.1972					
		SMSE	0.0527	0.0401					
	0.8	SM	1.0374	1.0520	0.7937	0.6281	0.4963	0.3883	0.3064
		SSE	0.2941	0.2824	0.0457	0.0770	0.1003	0.1118	0.1323
		ESE	0.2895	0.2610					
		SMSE	0.0879	0.0825					
	0.9	SM	1.0533	1.0768	0.8939	0.7978	0.7134	0.6368	0.5691
		SSE	0.3424	0.3787	0.0345	0.0643	0.0890	0.1072	0.1285
		ESE	0.3297	0.3000					
		SMSE	0.1201	0.1493					

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_2	0.5	SM	0.9591	1.3397	0.4988	0.2431	0.1092	0.0319	-0.0182
		SSE	0.2152	0.2829	0.0571	0.0727	0.0845	0.1030	0.1316
		ESE	0.2124	0.2385					
		SMSE	0.0480	0.1954					
	0.8	SM	0.9276	0.8751	0.8007	0.6392	0.5105	0.4054	0.3199
		SSE	0.2740	0.3790	0.0425	0.0758	0.1044	0.1295	0.1550
		ESE	0.2680	0.2567					
		SMSE	0.0803	0.1592					
	0.9	SM	0.9350	0.6313	0.8954	0.7995	0.7126	0.6365	0.5655
		SSE	0.2953	0.2770	0.0282	0.0533	0.0767	0.0957	0.1168
		ESE	0.2902	0.2449					
		SMSE	0.0914	0.2127					

T =6 NMP=0.95

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_1	0.5	SM	1.0207	1.0197	0.4915	0.2391	0.1151	0.0521	0.0231
		SSE	0.2225	0.1960	0.0549	0.0703	0.0768	0.0868	0.1066
		ESE	0.2133	0.1905					
		SMSE	0.0500	0.0388					
	0.8	SM	1.0457	1.0500	0.7930	0.6264	0.4943	0.3910	0.3080
		SSE	0.2983	0.2752	0.0430	0.0742	0.0939	0.1068	0.1237
		ESE	0.2827	0.2542					
		SMSE	0.0911	0.0782					
	0.9	SM	1.0558	1.0619	0.8943	0.7998	0.7144	0.6399	0.5746
		SSE	0.3392	0.3201	0.03347	0.0592	0.0806	0.1005	0.1172
		ESE	0.3162	0.2855					
		SMSE	0.1182	0.1063					

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_2	0.5	SM	0.6230	1.3170	0.4986	0.2445	0.1100	0.0323	-0.0153
		SSE	0.2177	0.2600	0.0503	0.0686	0.0784	0.0909	0.1209
		ESE	0.2069	0.2238					
		SMSE	0.0488	0.1681					
	0.8	SM	0.9466	0.9269	0.7967	0.6340	0.5048	0.4005	0.3145
		SSE	0.2817	0.3287	0.0388	0.0689	0.0918	0.1099	0.1308
		ESE	0.2653	0.2398					
		SMSE	0.0822	0.1134					
	0.9	SM	0.9292	0.6362	0.8976	0.8045	0.7220	0.6496	0.5843
		SSE	0.3130	0.2593	0.0270	0.0505	0.0723	0.0893	0.1052
		ESE	0.2892	0.2161					
		SMSE	0.1030	0.1996					

T=10 NMP=0.80

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_1	0.5	SM	1.0000	1.0281	0.4936	0.2398	0.1159	0.0574	0.0291
		SSE	0.6511	0.3171	0.0532	0.0685	0.0748	0.0818	0.0955
		ESE	0.2200	0.1957					
		SMSE	0.4239	0.1013					
	0.8	SM	1.0541	1.0628	0.7934	0.6276	0.4973	0.3936	0.3083
		SSE	0.3227	0.3116	0.0453	0.0788	0.0997	0.1181	0.1347
		ESE	0.3226	0.2882					
		SMSE	0.1071	0.1010					
	0.9	SM	1.1114	1.1031	0.8927	0.7934	0.7056	0.6255	0.5554
		SSE	0.4297	0.4004	0.0368	0.0752	0.1012	0.1260	0.1509
		ESE	0.4252	0.3379					
		SMSE	0.1976	0.1709					

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_2	0.5	SM	0.9198	1.8046	0.5113	0.2602	0.1283	0.0525	0.0030
		SSE	0.2105	0.4847	0.0528	0.0691	0.0784	0.0878	0.1058
		ESE	0.2125	0.3612					
		SMSE	0.0507	0.8823					
	0.8	SM	0.9087	1.4686	0.7908	0.6182	0.4789	0.3639	0.2660
		SSE	0.2820	0.6403	0.0391	0.0672	0.0894	0.1094	0.1311
		ESE	0.2811	0.4981					
		SMSE	0.0879	0.6296					
	0.9	SM	0.9340	1.3114	0.8815	0.7679	0.6644	0.5654	0.4742
		SSE	0.3959	0.6454	0.0294	0.0569	0.0822	0.1060	0.1308
		ESE	0.3192	0.8196					
		SMSE	0.1611	0.5133					

T=10 NMP=0.90

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_1	0.5	SM	1.0173	1.0120	0.4939	0.2446	0.1208	0.0569	0.0252
		SSE	0.2045	0.1698	0.0448	0.0569	0.0610	0.0648	0.0696
		ESE	0.1972	0.1726					
		SMSE	0.0421	0.0290					
	0.8	SM	1.0324	1.0423	0.7942	0.6307	0.5011	0.3976	0.3153
		SSE	0.2996	0.2540	0.0366	0.0622	0.0789	0.0915	0.1023
		ESE	0.2766	0.2437					
		SMSE	0.0908	0.0663					
	0.9	SM	1.0697	1.0607	0.8950	0.8003	0.7154	0.6407	0.5736
		SSE	0.3437	0.3179	0.0277	0.0528	0.0738	0.0911	0.1061
		ESE	0.3626	0.3226					
		SMSE	0.1230	0.1047					

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_2	0.5	SM	0.9433	1.5038	0.5101	0.2626	0.1339	0.0608	0.0155
		SSE	0.1842	0.2949	0.0424	0.0540	0.0610	0.0683	0.0767
		ESE	0.1908	0.2415					
		SMSE	0.0371	0.3408					
	0.8	SM	0.9212	1.1266	0.8012	0.6409	0.5116	0.4054	0.3184
		SSE	0.2447	0.4061	0.0322	0.0556	0.0749	0.0908	0.1054
		ESE	0.2533	0.2963					
		SMSE	0.0661	0.1809					
	0.9	SM	0.9191	0.7928	0.8966	0.8015	0.7155	0.6375	0.5663
		SSE	0.2868	0.3661	0.0218	0.0415	0.0615	0.0794	0.0963
		ESE	0.2850	0.3128					
		SMSE	0.0888	0.1770					

T=10 NMP=0.95

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_1	0.5	SM	1.0086	1.0164	0.4959	0.2448	0.1188	0.0547	0.0234
		SSE	0.1885	0.1708	0.0406	0.0493	0.0549	0.0582	0.0604
		ESE	0.1847	0.1603					
		SMSE	0.0356	0.0294					
	0.8	SM	1.0224	1.0234	0.7948	0.6310	0.5008	0.3970	0.3139
		SSE	0.2602	0.2372	0.0331	0.0546	0.0693	0.0809	0.0912
		ESE	0.2604	0.2271					
		SMSE	0.0682	0.0568					
	0.9	SM	1.0504	1.0490	0.8961	0.8024	0.7185	0.6438	0.5774
		SSE	0.3216	0.3039	0.0258	0.0483	0.0667	0.0815	0.0949
		ESE	0.3077	0.2692					
		SMSE	0.1060	0.0948					

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_2	0.5	SM	0.9415	1.4414	0.5075	0.2598	0.1328	0.0611	0.0180
		SSE	0.1736	0.2374	0.0390	0.0493	0.0547	0.0589	0.0647
		ESE	0.1775	0.2093					
		SMSE	0.0336	0.2512					
	0.8	SM	0.9370	1.0863	0.8003	0.6402	0.5115	0.4083	0.3248
		SSE	0.2373	0.3483	0.0296	0.0512	0.0680	0.0830	0.0968
		ESE	0.2431	0.2536					
		SMSE	0.0603	0.1288					
	0.9	SM	0.9342	0.7529	0.9009	0.8114	0.7298	0.6564	0.5901
		SSE	0.2708	0.3051	0.0206	0.0392	0.0567	0.0738	0.0906
		ESE	0.2759	0.2475					
		SMSE	0.0777	0.1541					

T=15 NMP=0.80

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_1	0.5	SM	1.0185	1.0188	0.4954	0.2449	0.1206	0.0554	0.0219
		SSE	0.2370	0.1979	0.0496	0.0657	0.0728	0.0778	0.0835
		ESE	0.2263	0.2010					
		SMSE	0.0565	0.0395					
	0.8	SM	1.0533	1.0634	0.7944	0.6308	0.5007	0.3976	0.3152
		SSE	0.3597	0.3237	0.0421	0.0744	0.0968	0.1121	0.1266
		ESE	0.3818	0.3517					
		SMSE	0.1322	0.1088					
	0.9	SM	1.0914	1.0780	0.8918	0.7941	0.7081	0.6310	0.5649
		SSE	0.4131	0.3923	0.0368	0.0712	0.0962	0.1198	0.1413
		ESE	0.4914	0.4449					
		SMSE	0.1790	0.1600					

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_2	0.5	SM	0.9052	2.2492	0.5260	0.2822	0.1529	0.0800	0.0328
		SSE	0.2431	0.6630	0.0486	0.0651	0.0732	0.0817	0.0921
		ESE	0.2315	0.6333					
		SMSE	0.0681	2.0000					
	0.8	SM	0.9246	1.983	0.7998	0.6381	0.5065	0.3975	0.3058
		SSE	0.2987	0.8125	0.0378	0.0642	0.0861	0.1042	0.1263
		ESE	0.3641	1.2519					
		SMSE	0.0949	1.6264					
	0.9	SM	0.9675	1.8141	0.8934	0.7965	0.7051	0.6192	0.5436
		SSE	0.3427	0.7990	0.0293	0.0551	0.0809	0.1069	0.1335
		ESE	0.4308	1.1240					
		SMSE	0.1185	1.3012					

T=15 NMP=0.90

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_1	0.5	SM	1.0082	1.0089	0.4954	0.2442	0.1202	0.0592	0.0266
		SSE	0.1785	0.1619	0.0405	0.0525	0.0540	0.0552	0.0556
		ESE	0.1715	0.1591					
		SMSE	0.0319	0.0263					
	0.8	SM	1.0282	1.0354	0.7931	0.6279	0.4961	0.3908	0.3077
		SSE	0.2586	0.2376	0.0327	0.0559	0.0722	0.0829	0.0904
		ESE	0.2511	0.2309					
		SMSE	0.0677	0.0577					
	0.9	SM	1.0532	1.0691	0.8939	0.7985	0.7132	0.6359	0.5673
		SSE	0.3238	0.3165	0.0269	0.0495	0.0684	0.0847	0.0999
		ESE	0.3197	0.2983					
		SMSE	0.1077	0.1049					

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_2	0.5	SM	0.9434	1.7202	0.5159	0.2705	0.1439	0.0758	0.0354
		SSE	0.1743	0.3427	0.0392	0.0497	0.0517	0.0533	0.0550
		ESE	0.1685	0.2592					
		SMSE	0.0336	0.6361					
	0.8	SM	0.9432	1.3943	0.8008	0.6401	0.5100	0.4040	0.3185
		SSE	0.2278	0.5175	0.0292	0.0500	0.0643	0.0769	0.0877
		ESE	0.2386	0.3544					
		SMSE	0.0551	0.4233					
	0.9	SM	0.9483	1.0860	0.8951	0.7989	0.7110	0.6318	0.5594
		SSE	0.2946	0.5602	0.0226	0.0420	0.0594	0.0750	0.0903
		ESE	0.2932	0.5045					
		SMSE	0.0895	0.3212					

T=15 NMP=0.95

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_1	0.5	SM	1.0073	1.0072	0.4977	0.2469	0.1208	0.0589	0.0288
		SSE	0.1636	0.1507	0.0374	0.0463	0.0474	0.0472	0.0475
		ESE	0.1606	0.1490					
		SMSE	0.0268	0.0228					
	0.8	SM	1.0234	1.0286	0.7941	0.6302	0.4995	0.3950	0.3128
		SSE	0.2475	0.2303	0.0272	0.0467	0.0587	0.0683	0.0759
		ESE	0.2366	0.2182					
		SMSE	0.0618	0.0539					
	0.9	SM	1.0424	1.0453	0.8960	0.8021	0.7182	0.6428	0.5756
		SSE	0.3064	0.3073	0.0219	0.0409	0.0557	0.0678	0.0787
		ESE	0.2894	0.2657					
		SMSE	0.0960	0.0965					

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_2	0.5	SM	0.9381	1.5373	0.5135	0.2687	0.1431	0.0773	0.0404
		SSE	0.1596	0.2479	0.0339	0.0436	0.0446	0.0455	0.0480
		ESE	0.1562	0.2037					
		SMSE	0.0293	0.3501					
	0.8	SM	0.9278	1.2621	0.8011	0.6416	0.5142	0.4111	0.3290
		SSE	0.2249	0.3804	0.0267	0.0448	0.0575	0.0680	0.0767
		ESE	0.2229	0.2696					
		SMSE	0.0558	0.2134					
	0.9	SM	0.9235	0.8933	0.9005	0.8112	0.7314	0.6590	0.5946
		SSE	0.2578	0.3604	0.0186	0.0338	0.0474	0.0602	0.0722
		ESE	0.2635	0.2855					
		SMSE	0.0723	0.1413					

Table A.3: **Non-Monotonic MCAR Case**: Simulated means (SM), simulated standard errors (SSE), simulated mean square error (SMSE), and estimated standard error (ESE) of the regression estimators based on GQL approach; SM and SSE of moment estimates for longitudinal correlation parameter under binary AR(1) process with T= 4, K=100, $\beta_1 = \beta_2 = 1$ and non-missing probabilities (NMP) 0.90 and 0.95 ; based on 1000 simulations.

T=4 NMP=0.90

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$
D_1	0.5	SM	1.0209	1.0321	0.4879	0.2490	0.1234
		SSE	0.2526	0.2326	0.0755	0.0923	0.1116
		ESE	0.2484	0.2220			
		SMSE	0.0642	0.0551			
	0.8	SM	1.0384	1.0559	0.7884	0.6294	0.5007
		SSE	0.3029	0.2934	0.0568	0.0928	0.1186
		ESE	0.3074	0.2766			
		SMSE	0.0932	0.0892			
	0.9	SM	1.0453	1.0619	0.8906	0.7989	0.7197
		SSE	0.3378	0.3285	0.0447	0.0780	0.1032
		ESE	0.3350	0.3025			
		SMSE	0.1162	0.1117			

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
D_2	0.5	SM	0.9746	1.2350	0.4874	0.2404	0.0952
		SSE	0.2446	0.2816	0.0664	0.0915	0.1180
		ESE	0.2390	0.2341			
		SMSE	0.0605	0.1345			
	0.8	SM	0.9543	0.8105	0.7981	0.6417	0.5133
		SSE	0.3012	0.3161	0.0511	0.0876	0.1151
		ESE	0.2835	0.2211			
		SMSE	0.0928	0.1358			
	0.9	SM	0.9776	0.6222	0.9008	0.8071	0.7209
		SSE	0.3333	0.2452	0.0349	0.0633	0.0858
		ESE	0.2997	0.2088			
		SMSE	0.1116	0.2029			

T=4 NMP=0.95

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
D_1	0.5	SM	1.0201	1.0320	0.4884	0.2438	0.1181
		SSE	0.2515	0.2312	0.0733	0.0903	0.1103
		ESE	0.2472	0.2207			
		SMSE	0.0637	0.0545			
	0.8	SM	1.0360	1.0551	0.7895	0.6268	0.4958
		SSE	0.3003	0.2915	0.0547	0.0909	0.1172
		ESE	0.3058	0.2748			
		SMSE	0.0915	0.0880			
	0.9	SM	1.0459	1.0636	0.8914	0.7979	0.7156
		SSE	0.3391	0.3195	0.0429	0.0757	0.1013
		ESE	0.3332	0.2990			
		SMSE	0.1171	0.1061			

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
D_2	0.5	SM	0.9740	1.2287	0.4880	0.2355	0.0912
		SSE	0.2434	0.2747	0.0649	0.0894	0.1158
		ESE	0.2379	0.2326			
		SMSE	0.0599	0.1278			
	0.8	SM	0.9486	0.8103	0.7979	0.6390	0.5087
		SSE	0.2993	0.3057	0.0500	0.0866	0.1147
		ESE	0.2825	0.2205			
		SMSE	0.0922	0.1294			
	0.9	SM	0.9483	0.6014	0.9006	0.8073	0.7214
		SSE	0.3062	0.2153	0.0341	0.0628	0.0843
		ESE	0.2958	0.1947			
		SMSE	0.0964	0.2052			

Table A.4: **Monotonic MAR Models 1 and 2**: Simulated means (SM), simulated standard errors (SSE), simulated mean square error (SMSE), and estimated standard error (ESE) of the regression estimators based on GQL approach; SM and SSE of moment estimates for longitudinal correlation parameter under binary AR(1) process with T= 6, K=100, $\beta_1 = \beta_2 = 1$; based on 1000 simulations.

`T=6 MODEL:M1`

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_1	0.5	SM	1.0371	1.0294	0.4903	0.2394	0.1109	0.0464	0.0196
		SSE	0.2815	0.2473	0.0769	0.1165	0.1437	0.1822	0.2616
		ESE	0.2702	0.2396					
		SMSE	0.0806	0.0620					
	0.8	SM	1.0743	1.0793	0.7907	0.6171	0.4836	0.3778	0.3031
		SSE	0.3695	0.3650	0.0625	0.1287	0.1757	0.2227	0.2729
		ESE	0.3755	0.3403					
		SMSE	0.1421	0.1395					
	0.9	SM	1.0934	1.0706	0.8909	0.7864	0.7020	0.6274	0.5612
		SSE	0.4143	0.3674	0.0515	0.1228	0.1805	0.2393	0.3005
		ESE	0.4322	0.3791					
		SMSE	0.1804	0.1400					

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_2	0.5	SM	0.9254	1.9805	0.4828	0.2042	0.0366	-0.1067	-0.2591
		SSE	0.2548	0.4511	0.0716	0.0965	0.1253	0.1708	0.2503
		ESE	0.2538	0.4284					
		SMSE	0.0705	1.1649					
	0.8	SM	0.9416	1.9273	0.7419	0.5135	0.3154	0.1320	-0.0312
		SSE	0.3163	0.6348	0.0627	0.1106	0.1574	0.2137	0.2850
		ESE	0.3041	0.5340					
		SMSE	0.1035	1.2629					
	0.9	SM	0.9578	1.8289	0.8287	0.6498	0.4763	0.3138	0.1732
		SSE	0.3442	0.7482	0.0606	0.1167	0.1738	0.2332	0.3072
		ESE	0.3298	0.5976					
		SMSE	0.1203	1.2469					

T=6 MODEL:M2

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_1	0.5	SM	1.0451	1.0277	0.4913	0.2381	0.1181	0.0526	0.0257
		SSE	0.2925	0.2591	0.0753	0.1090	0.1338	0.1760	0.2399
		ESE	0.2709	0.2402					
		SMSE	0.0876	0.0679					
	0.8	SM	1.0779	1.0598	0.7858	0.6137	0.4820	0.3813	0.3190
		SSE	0.3752	0.3142	0.0632	0.1193	0.1713	0.2270	0.2875
		ESE	0.3743	0.3231					
		SMSE	0.1468	0.1023					
	0.9	SM	1.0737	1.0699	0.8900	0.7875	0.7014	0.6311	0.5808
		SSE	0.4241	0.3752	0.0514	0.1163	0.1739	0.2324	0.2945
		ESE	0.4447	0.4019					
		SMSE	0.1853	0.1457					

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$	$\hat{\rho_4}$	$\hat{\rho_5}$
D_2	0.5	SM	0.9142	1.9901	0.4767	0.1931	0.0132	-0.1266	-0.2129
		SSE	0.2454	0.4790	0.0721	0.1006	0.1316	0.1785	0.2667
		ESE	0.2481	0.4141					
		SMSE	0.0676	1.2097					
	0.8	SM	0.9351	1.9914	0.7311	0.4930	0.2828	0.0999	-0.0351
		SSE	0.2983	0.6483	0.0660	0.1139	0.1640	0.2176	0.2814
		ESE	0.2990	0.5245					
		SMSE	0.0932	1.4032					
	0.9	SM	0.9685	1.9408	0.8151	0.6265	0.4528	0.2948	0.1706
		SSE	0.3416	0.7432	0.0681	0.1322	0.1878	0.2484	0.3163
		ESE	0.3225	0.6177					
		SMSE	0.1177	1.4375					

Table A.5: **Non-Monotonic MAR Models 1 and 2**: Simulated means (SM), simulated standard errors (SSE), simulated mean square error (SMSE), and estimated standard error (ESE) of the regression estimators based on GQL approach; SM and SSE of moment estimates for longitudinal correlation parameter under binary AR(1) process with T= 4, K=100, $\beta_1 = \beta_2 = 1$; based on 1000 simulations.

MODEL: M1

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$
D_1	0.5	SM	1.0340	1.0108	0.4927	0.2839	0.1528
		SSE	0.2694	0.2476	0.0873	0.1114	0.1485
		ESE	0.2731	0.2419			
		SMSE	0.0737	0.0614			
	0.8	SM	1.0614	1.0306	0.7913	0.6531	0.5366
		SSE	0.3625	0.3200	0.0711	0.1103	0.1549
		ESE	0.3804	0.3323			
		SMSE	0.1352	0.1033			
	0.9	SM	1.0606	1.0337	0.8918	0.8132	0.7444
		SSE	0.3718	0.3464	0.0621	0.0970	0.1451
		ESE	0.4037	0.3859			
		SMSE	0.1419	0.1211			

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$
D_2	0.5	SM	0.9545	1.3892	0.4696	0.2447	0.0799
		SSE	0.2620	0.3581	0.0772	0.1099	0.1543
		ESE	0.2520	0.2629			
		SMSE	0.0707	0.2797			
	0.8	SM	0.9727	1.0800	0.7606	0.6066	0.4702
		SSE	0.3205	0.4603	0.0681	0.1154	0.1678
		ESE	0.2938	0.3105			
		SMSE	0.1035	0.2183			
	0.9	SM	0.9910	0.9000	0.8681	0.7648	0.6687
		SSE	0.3660	0.4481	0.0559	0.0983	0.1427
		ESE	0.3677	0.2661			
		SMSE	0.1340	0.2108			

MODEL:M2

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$
D_1	0.5	SM	1.0353	1.0117	0.4907	0.2856	0.1510
		SSE	0.2661	0.2470	0.0857	0.1132	0.1450
		ESE	0.2731	0.2421			
		SMSE	0.0721	0.0611			
	0.8	SM	1.0508	1.0237	0.7904	0.6512	0.5365
		SSE	0.3372	0.3032	0.0694	0.1082	0.1557
		ESE	0.3422	0.3027			
		SMSE	0.1163	0.0925			
	0.9	SM	1.0612	1.0365	0.8913	0.8126	0.7432
		SSE	0.3744	0.3464	0.0630	0.0966	0.1480
		ESE	0.4020	0.3564			
		SMSE	0.1439	0.1213			

Design	ρ	Statistic	$\hat{\beta_{1G}}$	$\hat{\beta_{2G}}$	$\hat{\rho_1}$	$\hat{\rho_2}$	$\hat{\rho_3}$
D_2	0.5	SM	0.9540	1.4301	0.4557	0.2393	0.0853
		SSE	0.2609	0.3502	0.0790	0.1117	0.1531
		ESE	0.2508	0.2629			
		SMSE	0.0702	0.3076			
	0.8	SM	0.9735	1.1008	0.7545	0.6053	0.4741
		SSE	0.3167	0.4601	0.0724	0.1181	0.1647
		ESE	0.2927	0.2694			
		SMSE	0.1010	0.2219			
	0.9	SM	1.0024	0.9231	0.8640	0.7606	0.6674
		SSE	0.3961	0.4487	0.0579	0.0995	0.1382
		ESE	0.3504	0.2818			
		SMSE	0.1569	0.2072			

Bibliography

[1] Binder, D. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review* **51**, 279–292.

[2] Crowder, M. (1995). On the use of a Working Correlation Matrix in Using Generalized Linear Models for repeated measures. *Biometrika* **82**, 407–410.

[3] Jowaheer, V. , and Sutradhar, B. C (2002). Analysing longitudinal count data with overdispersion. *Biometrika* **89 , 2**, 389–399.

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

[5] Little, R. J. A. , and Rubin, D. B. (1987). *Statistical Analysis with Missing Data. New York: John Wiley.*

[6] McCullagh , P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11**, 59–67.

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

[8] Robins , J. M. , and Rotnitzky , A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, **90**, 122–129.

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

[10] Sutradhar , B. C., and Das , K. (1999). On the efficiency of regression estimators in generalized linear models for longitudinal data. *Biometrika* **86**, 459–465.

[11] Sutradhar , B. C., and Kovacevic, M. (2003). Analysing longitudinal survey data in the presence of missing responses. *Technical Report, Department of Mathematics and Statistics, Memorial University of Newfoundland.*

[12] Xie , F., and Paik ,M. C. (1997). Generalized estimating equation model for binary outcomes with missing covariates. *Biometrika* **53**, 1458–1466.