Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 64(2) 2010 177-189

Inferences in Longitudinal Mixed Models for Survey Data

Brajendra Sutradhar[1*], R. Prabhakar Rao[2] and V.N. Pandit[2]
[1]*Department of Mathematics and Statistics, Memorial University, Canada*
[2]*Department of Economics, Sri Sathya Sai University, India*

Received 10 March 2010; Accepted 09 May 2010

SUMMARY

In sample survey based longitudinal set up, first, a suitable sample of individuals are chosen from a finite or survey population by using an appropriate sampling technique such as two stage cluster sampling. Secondly, a response of interest along with a set of multi-dimensional covariates are collected from each individual of the sample, over a small period of time. These repeated responses exhibit longitudinal correlations. Also, on top of the influence of time dependent covariates, the responses of the individual may further be influenced by an individual random effect. Since, the so-called generalized quasi-likelihood (GQL) approach has been shown to be an efficient estimation approach for both regression effects and random effects variance involved in an infinite population based longitudinal mixed model (Sutradhar *et al*. 2008), for example), in this paper we demonstrate how to develop the sample survey based GQL estimating equations for the estimation of the desired finite population parameters both in linear dynamic and binary dynamic mixed models set up. We also illustrate the sampling design weights based GQL estimation methodology by re-analyzing the SLID (Survey of Labor and Income Dynamic) data from Statistics Canada.

Keywords: Consistency, Dynamic dependence parameters, Efficiency, Random effects, Regression effects, Sampling design weights, Variance components.

1. INTRODUCTION

Panel data analysis is an important research problem in socio-economic and biomedical fields, among others. In this set up, it is common to collect a small number of repeated continuous or discrete such as binary responses along with multi-dimensional covariates from a group of selected individuals. For example, in a provincial clinical trial study, it may be of interest to know the effects of certain covariates such as gender, age, smoking status and food habits on the blood pressures of an individual collected over a period of a few weeks. In this example, blood pressure is assumed to have a linear relationship with the covariates, which may also be affected by individual random effects. Similarly, in a provincial economics study, a state government may be interested to know the effects of certain covariates such as gender, age, education level and marital status on the employment status of an individual collected over a small period of time. In this second problem, employment status is a binary variable, which may be assumed to have a logistic relationship to the covaraites, and the employment status may also be affected by individual random effects. Note that since the state has a huge number of individuals to consider for each of the above two problems, it would be appropriate to choose a small number of individual in an efficient way, and then collect necessary information from the selected individuals over a period of time. The scientific question is to analyze the longitudinal data from the selected individuals in order to ascertain the effects of the covariates on the responses by taking the variation of the random individual effects into account.

[*]*Corresponding author* : Brajendra Sutradhar
E-mail address : bsutradh@mun.ca

Now suppose that there are N individuals in a survey population and y_{it} denotes the continuous or discrete response collected from the i-th (i = 1, ..., N) at time point t over a duration t = 1, ..., T. Let $x_{it} = (x_{it1}, ..., x_{itp})'$ denote the p–dimensional time dependent covariate vector corresponding to y_{it}. Next, we assume that the responses y_{it} and the corresponding covariate vector x_{it} in the survey population follow a super population model (M) to explain the effect of x_{it} on y_{it}. For definitions of the survey population and super-population one may refer to Godambe and Thompson (1986), for example.

For continuous data, we consider a linear dynamic mixed (LDM) super population model given by

$$y_{i1} = x_{i1}'\beta + z_i\gamma_i^* + \varepsilon_{i1}$$

$$y_{it} = x_{it}'\beta + \theta(y_{i,t-1} - x_{i,t-1}'\beta) + z_i\gamma_i^* + \varepsilon_{it} \quad \text{for } t = 2, ..., T \tag{1.1}$$

[see Amemiya (1985, Section 6.6.3), Hsiao (2003), Arellano and Bond (1991), Imbens (2002), Chamberlain (1992), and Bun and Carree (2005)] where γ_i^* is an unobservable random effect of the i^{th} individual, ε_{it} is error component of the linear model, and z_i is a known additional covariate for the i^{th} individual on top of the fixed covariates x_{it}. In (1.1), β is the $p \times 1$ vector of fixed effects of x_{it} on y_{it}, and θ is a scalar dynamic dependence effect of $y_{i,t-1}$ on y_{it} for t = 2, ..., T. Here both β and θ are referred to as the super population parameter. Further we assume that

$$\varepsilon_{it} \overset{iid}{\sim} (0, \sigma_\varepsilon^2) \text{ and } \gamma_i^* \overset{iid}{\sim} (0, \sigma_\gamma^2) \tag{1.2}$$

where σ_ε^2 is the error variance and σ_γ^2 is the random effects variance. Similar to β and θ, these variances, namely, σ_ε^2 and σ_γ^2 are also super population parameters. Also, we assume that ε_{it} and γ_i are independent. Now it follows from (1.1)-(1.2) that the super population model (M) based first order moment is given by

$$E_M[Y_{it}] = x_{it}'\beta = \mu_{it} \text{ (say)} \tag{1.3}$$

which is a function of the super population parameter β only, whereas the model based second order moments given by

$$\text{var}_M[Y_{it}] = \sigma_{itt} = z_i^2\sigma_\gamma^2 \left\{ \sum_{j=0}^{t-1} \theta^j \right\}^2 + \sigma_\varepsilon^2 \sum_{j=0}^{t-1} \theta^{2j}$$

$$\text{for } t = 1, ..., T \tag{1.4}$$

and

$$\text{cov}_M[Y_{iu}, Y_{it}] = \sigma_{iut} = z_i^2\sigma_\gamma^2 \sum_{j=0}^{t-1} \theta^j \sum_{k=0}^{u-1} \theta_k + \sigma_\varepsilon^2 \sum_{j=0}^{u-1} \theta^{t-u+2j}, \text{ for } u < t \tag{1.5}$$

are functions of the super population parameter θ, σ_γ^2, and σ_ε^2.

For longitudinal binary data, Sutradhar *et al*. (2008) have considered a non-linear binary dynamic logistic mixed (NLDLM) super population model given by

$$Pr(y_{it} = 1|\gamma_i) = \begin{cases} \dfrac{\exp(x_{i1}'\beta + \sigma_\gamma\gamma_i)}{1 + \exp(x_{i1}'\beta + \sigma_\gamma\gamma_i)} & \text{for } t = 1 \\ \dfrac{\exp(x_{it}'\beta + \theta y_{i,t-1} + \sigma_\gamma\gamma_i)}{1 + \exp(x_{it}'\beta + \theta y_{i,t-1} + \sigma_\gamma\gamma_i)} & \text{for } t = 2, ..., T \end{cases}$$

where $\gamma_i \overset{iid}{\sim} (0, 1)$. This model produces marginal means those maintain a recursive relation among them. By the same token, the variances also maintain a recursive relation. As far as the correlations are concerned, this model produces longitudinal correlations ranging from –1 to +1.

In some situations, it may be, however, more appropriate to assume that the marginal mean at a given point of time depends only on the covariates at that time, and thus it does not have any relationship with means from the past. One may use a conditionally linear

dynamic probability model to represent such a situation. This conditional linear model may be written as

$$\Pr(y_{it} = 1|\gamma_{i,y_{i,t-1}}, ..., y_{i1}) = \mu^*_{it}(\gamma_i) + \sum_{j=1}^{t-1} \theta^*_{i,tj}(\gamma_i) (y_{i,t-j} - \mu^*_{i,t-j}(\gamma_i)) \quad (1.6)$$

(see Qaqish (2003) for a similar but fixed effects model) where

$$\mu^*_{it}(\gamma_i) = \frac{\exp(x'_{it}\beta + \sigma_\gamma \gamma_i)}{1 + \exp(x'_{it}\beta + \sigma_\gamma \gamma_i)} \quad (1.7)$$

yielding $\sigma^*_{itt}(\gamma_i) = \mu^*_{it}(\gamma_i)\,[1 - \mu^*_{it}(\gamma_i)]$. One may then compute the first order unconditional moment by using

$$\begin{aligned} \mu_{it} &= E_M[Y_{it}] = E_{\gamma i} E_M[Y_{it}|\gamma_i] \\ &= E_{\gamma i}\,[\mu^*_{it}(\gamma_i)] \\ &= W^{-1} \sum_{w=1}^{W} \mu^*_{it}(\gamma_{iw}) \end{aligned} \quad (1.8)$$

[Jiang (1998), Sutradhar (2004)] where γ_{iw} is the w^{th} ($w = 1, ..., W$) realized value of γ_i generated from the standard normal distribution. Here W is a sufficiently large number, such as $W = 5000$. Consequently, the model based variance of y_{it} is computed as

$$\begin{aligned} \sigma_{itt} &= \text{var}_M[Y_{it}] \\ &= \mu_{it}[1 - \mu_{it}] \\ &= W^{-1} \sum_{w=1}^{W} \mu^*_{it}(\gamma_{iw})\ [1 - W^{-1} \sum_{w=1}^{W} \mu^*_{it}(\gamma_{iw})] \end{aligned} \quad (1.9)$$

Next following (1.6), one may relate y_{it} and y_{iu} as

$$E[Y_{it}|\gamma_i, y_{i,t-1}, ... y_{iu}) = \mu^*_{it}(\gamma_i) + \sum_{j=1}^{t-u} \theta^*_{i,tj}(\gamma_i) (y_{i,t-j} - \mu^*_{i,t-j}(\gamma_i)) \quad (1.10)$$

and compute the model based conditional covariance by using

$$\begin{aligned} \sigma^*_{iut}(\gamma_i) &= \text{cov}_M\{(Y_{iu}, Y_{it})|\gamma_i\} \\ &= E_M[(Y_{iu} - \mu^*_{iu}(\gamma_i))(Y_{it} - \mu^*_{it}(\gamma_i))] \\ &= \Big[\Pi^i_{j=1}\theta^*_{i,t-j+1,1}(\gamma_i) + \Pi^{u-1}_{j=1}\theta^*_{i,t-j+1,2}(\gamma_i) \\ &\quad + ... + \Pi^1_{j=1}\theta^*_{i,t-j+1,t-u}(\gamma_i)]\,\sigma^*_{iuu}(\gamma_i) \end{aligned} \quad (1.11)$$

with $\sigma^*_{iuu}(\gamma_i) = \mu^*_{iu}\,[1 - \mu^*_{iu}]$, Note that at a given point of time t

$$\theta^*_{it}(\gamma_i) = \Big[\theta^*_{t,t1}(\gamma_i),...,\theta^*_{i,t,t-1}(\gamma_i)\Big]'$$

may be computed from a pre-specified correlation structure by using the formula

$$\theta^*_{it}(\gamma_i) = \Big\{Cov(y^*_{i,t-1}|\gamma_i)\Big\}^{-1}\Big[Cov(y^*_{i,t-1}, y_{it})|\gamma_i\Big] \quad (1.12)$$

where $y^*_{i,t-1} = (y_{i1}, ... y_{i,t-1})'$

$$Cov(y^*_{i,t-1}) = (A^*_i)^{1/2}\, C_i(\rho)(A^*_i)^{1/2} \quad (1.13)$$

with $C_i(\rho) = (c_{i,jt})$ as the pre-specified correlation structure, and

$$A^*_i = \text{diag}\Big[a^*_{i,11}, ..., a^*_{i,t-1,t-1}\Big]$$

where $a^*_{i,jj} = \sigma^*_{ijj}(\gamma_i) = \mu^*_{ij}(\gamma_i)(1 - \mu^*_{ij}(\gamma_i))$

Furthermore,

$$Cov(y^*_{i,t-1}, y_{it}) = \Big(\sqrt{a^*_{i,tt}a^*_{i,11}}\,c_{i,1t}, ..., \sqrt{a^*_{i,tt}a^*_{i,t-1,t-1}}c_{i,t-1,t}\Big)' \quad (1.14)$$

As far as the range for the correlation index parameter ρ is concerned, this can be computed by using the fact that the conditional probability in (1.6) has to be bounded between 0 and 1, that is $0 < \Pr[y_{it} = 1|\gamma_i, y_{i,t-1} ..., y_{i1}] < 1$, for all $i = 1, ..., K$.

Note that when it is of interest to compute the regression effects β only, one may compute the

conditional covariance $\text{cov}[(Y_{iu}, Y_{it})|\gamma_i]$ in (1.11) by using an equivalent formula

$$\text{cov}[(Y_{iu}, Y_{it})|\gamma_i] = \rho_{|t-u|}\sqrt{[\text{var}(Y_{iu}|\gamma_i)\ \text{var}(Y_{it}|\gamma_i)]}$$
$$= \rho_{|t-u|}\sqrt{[\sigma^*_{iuu}(\gamma_i)\,\sigma^*_{itt}(\gamma_i)]} \quad (1.15)$$

where $\rho_{|t-u|}$ is the $(u, t)^{\text{th}}$ element of the $C_i(\rho)$ matrix and σ^*_{itt}, for example, is the conditional variance given by $\sigma^*_{itt} = \mu^*_{it}[1 - \mu^*_{it}]$. It then follows that the model based unconditional covariance $\text{cov}_M[Y_{iu}, Y_{it}]$ may be computed by using the formula

$$\sigma_{iut} = \text{cov}_M[Y_{iu}, Y_{it}]$$
$$= \rho_{|t-u|}E_{\gamma i}\left[\sqrt{[\sigma^*_{iuu}(\gamma_i)\,\sigma^*_{itt}(\gamma_i)]}\right] + \text{cov}_{\gamma i}[E\{Y_{iu}\mid\gamma_i\}, E\{Y_{it}\mid\gamma_i\}]$$
$$= \rho_{|t-u|}W^{-1}\sum_{w=1}^{W}\left[\sqrt{[\sigma^*_{iuu}(\gamma_{iw})\,\sigma^*_{itt}(\gamma_{iw})]}\right]$$
$$+W^{-1}\sum_{w=1}^{W}[\mu^*_{iu}(\gamma_{iw})\mu^*_{it}(\gamma_{iw})] - \mu_{iu}\mu_{it} \quad (1.16)$$

with μ_{it} as in (1.8), leading to the model based uncorrected second order product moments as

$$\lambda_{iut} = E_M[Y_{iu}Y_{it}] = \sigma_{iut} + \mu_{iu}\mu_{it} \quad (1.17)$$

Note that the aforementioned model based first and second order moments, namely the formulas for μ_{it} and σ_{iut} given, respectively, in (1.3) and (1.4)-(1.5) under the linear dynamic mixed model, and in (1.8) and (1.16)-(1.17), respectively, under the dynamic binary mixed model, will be exploited in Section 2 to develop the survey population (finite) based generalized quasi-likelihood (GQL) estimating equations for β. Similarly, by exploiting the model based third and fourth order moments, the survey population based GQL estimating equations for θ, σ^2_γ, and σ^2_ε under the linear dynamic mixed model, and for σ^2_γ under the dynamic binary mixed model, are developed in Section 2. The longitudinal correlation parameters such as ρ under the binary dynamic mixed model will be computed by using the well known method of moments (MM) which always produces consistent estimates.

Further note that since N, the total number of individuals in the finite or survey population, is quite large, it is practical to assume that the responses from all N individuals in the finite population are not available. Consequently, one can not use the survey population based GQL estimating equations to be developed in Section 2 to estimate the finite population parameters. As a remedy, it is natural to consider an appropriate sample s^* of size n individuals from N individuals of the finite population. As far as the sampling technique is concerned, we will use a two-stage clustered sampling, for example. The construction of the sampling design weights (SDW) (Binder 1983, Pfeffermann 1993) based GQL estimating equations is shown in Section 3, for the estimation of the survey population parameters defined through the survey population based GQL estimating equations from Section 2. Both continuous and binary cases are considered in developing such SDW based weighted GQL (WGQL) estimating equations. The WGQL approach is illustrated by re-analyzing the well known SLID (survey of labor and income dynamics) data from Statistics Canada. This data set contains longitudinal binary responses on employment/unemployment along with certain suitable covariates collected from 15,731 individuals selected by using a two-stage complex sampling design. The effects of ignoring the SDW in estimating the survey population parameters is also examined. Note that Sutradhar and Kovacevic (2000) have earlier analyzed the SLID data for two years 1993-1994, but, by using an ordinal longitudinal model, whereas in this paper we use repeated binary responses for six years from 1993 to 1998.

2. FINITE POPULATION BASED GQL ESTIMATING EQUATIONS

2.1 Finite Population Based Estimation of β

Recall that the super population means (μ_{it}), variances (σ_{itt}) and covariances (σ_{iut}) of the repeated responses $y_{i1}, ..., y_{it}, ..., y_{iT}$ are given by (1.3) and (1.4)-(1.5), respectively, under the linear mixed model (1.1), and by (1.8) and (1.16), respectively, under the non-linear binary mixed model (1.6). Let $y_i = (y_{i1}, ..., y_{it}, ...; y_{iT})'$ be the vector of T responses for the i^{th} individual, so that,

$$\mu_i = E_M(Y_i) = (\mu_{i1}, ..., \mu_{it}, ..., \mu_{iT})', \text{ and}$$
$$\Sigma_i = \text{cov}_M(Y_i) = (\sigma_{iut}) \quad (2.1)$$

Now if the responses along with covariate information were available from all N individuals of the survey population, then following Sutradhar (2004), one could obtain the finite population (survey population) based consistent and efficient estimates for the super population parameter β by solving the survey population (finite) based GQL estimating equation given by

$$\sum_{i=1}^{N} \frac{\partial \mu_i'(\beta)}{\partial \beta} \sum_i^{-1} (y_i - \mu_i(\beta)) = \sum_{i=1}^{N} g_{i1}(\beta|\theta, \sigma_\gamma^2, \sigma_\in^2) = G_1(\beta|\theta, \sigma_\gamma^2, \sigma_\in^2) = 0 \quad (2.2)$$

in the linear dynamic mixed model case, and by solving the survey population (finite) based GQL estimating equation given by

$$\sum_{i=1}^{N} \frac{\partial \mu_i'(\beta)}{\partial \beta} \sum_i^{-1} (y_i - \mu_i(\beta)) = \sum_{i=1}^{N} g_{i1}^*(\beta|\theta, \sigma_\gamma^2) = G_1^*(\beta|\theta, \sigma_\gamma^2) = 0 \quad (2.3)$$

in the binary dynamic mixed model case.

Let β_N denote this finite population based GQL estimate of β. Note that this estimate β_N is also referred to as the survey population parameter.

2.2 Finite Population Based Estimation of $\xi = \left(\theta, \sigma_\gamma^2, \sigma_\in^2\right)'$ for Linear Dynamic Mixed Model

Note that θ is a dynamic dependence parameter and σ_γ^2 is the variance of the random effects. In a non-linear dynamic mixed model set up for binary data, Sutradhar *et al*. (2008) used the raw second order response based GQL approach for the estimation of this type of dynamic dependence and variance component parameters. To develop similar GQL approach in the linear mixed model set up, it is convenient to use the corrected second order responses and construct a vector of basic statistic given by

$$s_i = [(y_{i1} - \mu_{i1})^2, ..., (y_{iT} - \mu_{iT})^2, (y_{i1} - \mu_{i1})(y_{i2} - \mu_{i2}), ..., (y_{iu} - \mu_{iu})(y_{it} - \mu_{it}), ..., (y_{i,\,T-1} - \mu_{i,\,T-1})(y_{iT} - \mu_{iT})]' \quad (2.4)$$

The super population model based expectation of s_i may then be written as

$$\sigma_i = E_M(s_i) = (\sigma_{i11}, ..., \sigma_{itt}, ..., \sigma_{iTT}, \sigma_{i12}, ..., \sigma_{iut}, ..., \sigma_{i,\,T-1,T})' \quad (2.5)$$

where σ_{itt} and σ_{iut} are given by (1.4)-(1.5), and they are function of θ, σ_γ^2, and $\sigma_\in^2$ parameters. Let the super population model based covariance matrix of s_i be denoted by Ω_i, that is, $\Omega_i = \text{cov}_M(s_i)$. Now if the responses and covariates for the computation of s_i were available from all N individuals of the survey population, then following Sutradhar *et al*. (2008), one could obtain the finite population (survey population) based consistent and efficient estimates for the super population parameter $\xi = (\theta, \sigma_\gamma^2, \sigma_\in^2)$ by solving the survey population (finite) based GQL estimating equation given

$$\sum_{i=1}^{N} \frac{\partial \sigma_i'}{\partial \xi} \Omega_i^{-1} (s_i - \sigma_i) = \sum_{i=1}^{N} g_{i2}(\theta, \sigma_\gamma^2, \sigma_\in^2|\beta) = G_2(\theta, \sigma_\gamma^2, \sigma_\in^2|\beta) = 0 \quad (2.6)$$

which is however not possible to compute under the present linear dynamic mixed model (1.1). This is because, the distributions of the random effects γ_i^* and of $\in_{it}$ under (1.1) are not known and hence the one can not compute the fourth order moments based Ω_i matrix. However, since the consistent estimation of ξ is not affected by the choice of the weight matrix Ω_i, we, for convenience, use normality (N^*) based weight matrix Ω_{i,N^*}, that is computed by pretending that the responses follow the multivariate normal distribution. To reflect this normal approximation, one may re-write the GQL estimating equations in (2.6) as

$$\sum_{i=1}^{N} \frac{\partial \sigma_i'}{\partial \xi} \Omega_{iN^*}^{-1} (s_i - \sigma_i) = \sum_{i=1}^{N} g_{i2,N^*}(\theta, \sigma_\gamma^2, \sigma_\in^2|\beta) = G_{2,N^*}(\theta, \sigma_\gamma^2, \sigma_\in^2|\beta) = 0 \quad (2.7)$$

where Ω_{iN^*} may be computed by using the normality based third and fourth order moments given by following two lemmas.

Lemma 2.1 Let $\delta^*_{iult} = E_{M|N^*}[(Y_{iu} - \mu_{iu})(Y_{il} - \mu_{il})(Y_{it} - \mu_{it})]$, yielding

$$\delta^*_{iult} = 0 \tag{2.8}$$

Lemma 2.2 Let $\phi^*_{iulmt} = E_{M|N^*}[(Y_{iu} - \mu_{iu})(Y_{il} - \mu_{il})(Y_{im} - \mu_{im})(Y_{it} - \mu_{it})]$. It then follows that

$$\phi^*_{iulmt} = \sigma_{iul}\sigma_{imt} + \sigma_{ium}\sigma_{ilt} + \sigma_{iut}\sigma_{ilm} \tag{2.9}$$

where σ_{iut}, for example, is available from (1.5).

Let $\xi_N = (\theta_N, \sigma^2_{\gamma,N}, \sigma^2_{\in,N})'$ be the finite population based estimate of the super population parameter $\xi = (\theta, \sigma^2_\gamma, \sigma^2_\in)'$, obtained from (2.7).

2.3 Finite Population Based Estimation of $\alpha = (\sigma^2_\gamma, \rho_1, \ldots, \rho_{T-1})'$ for Binary Dynamic Mixed Model

2.3.1 GQL estimation of σ^2_γ

To construct the GQL estimating equations for σ^2_γ for the binary data in the infinite population set up, Sutradhar *et al*. (2008) used a raw second order response vector

$$u_i = (u'_{i1}, u'_{i2})' \tag{2.10}$$

where

$u_{i1} = (y^2_{i1}, \ldots, y^2_{iT})' : T \times 1$

$u_{i2} = (y_{i1}y_{i2}, \ldots, y_{iu}y_{it}, \ldots, y_{i,T-1}y_{iT})' : (T-1)T/2 \times 1$

Note that since $y^2_{it} \equiv y_{it}$ for the binary data, one may write

$$y_i = u_{i1} = (y_{i1}, \ldots, y_{it}, \ldots, y_{iT})'$$

and hence u_i in (2.10) can be re-expressed as

$$u_i = (y'_i, s'_i)' \tag{2.11}$$

with $s_i = u_{i2}$, for convenience of notation. Let

$$\eta_i = E_M(U_i) = [E(Y'_i), E(S'_i)]' \tag{2.12}$$

be the super population model based expectation of the vector u_i, which is computed as

$$\eta_i = [\mu'_i, \lambda'_i]' \tag{2.13}$$

where

$$\mu_i = [\mu_{i1}, \ldots, \mu_{it}, \ldots, \mu_{iT}]'$$

$$\lambda_i = [\lambda_{i12}, \ldots, \lambda_{iut}, \ldots, \lambda_{i,T-1,T}]'$$

with $\mu_{it} = E[Y_{it}]$ and $\lambda_{iut} = E[Y_{iu}Y_{it}]$ as given in (1.8) and (1.17), respectively.

Further, let Γ_i be the super population model based covariance matrix of u_i. For convenience, we express this matrix as

$$\Gamma_i = \begin{pmatrix} \Sigma_i & \Delta_i \\ \Delta'_i & \Phi_i \end{pmatrix} = \begin{pmatrix} \mathrm{cov}_M(Y_i) & \mathrm{cov}_M(Y_i, S'_i) \\ \mathrm{cov}_M(S_i, Y'_i) & \mathrm{cov}_M(S_i) \end{pmatrix} \tag{2.14}$$

where $\Sigma_i = (\sigma_{iut})$ with σ_{itt} and σ_{iut} $(u < t)$ as given by (1.9) and (1.16), respectively. Note that unlike the computation for Σ_i, one can not compute the third and fourth order moments under the dynamic binary mixed model (1.6). One may however obtain an approximation to these moments by pretending that the binary response vector $y_i = [y_{i1}, \ldots, y_{iT}]'$ follows the T-dimensional normal distribution but with true binary mean vector $\mu_i = [\mu_{i1}, \ldots, \mu_{iT}]'$ defined by (1.8), and true binary covariance matrix Σ_i defined by (1.9) and (1.16). We refer to Prentice and Zhao (1991), among others, for such an approximation. Let this normal (N^*) approximation yields Δ_{i,N^*} and Φ_{i,N^*}, for Δ_i and Φ_i, respectively.

Now, the elements of Δ_{i,N^*} matrix may be computed by using the general formula

$$\begin{aligned} \delta_{iult,N^*} &= \mathrm{cov}_M[Y_{iu}, Y_{il}Y_{it}] \\ &= E_M[Y_{iu}Y_{il}Y_{it}] - \mu_{iu}\lambda_{ilt} \end{aligned} \tag{2.15}$$

where, following the Lemma 2.1, the third order raw moments are computed as

$$\begin{aligned} E_M[Y_{iu}Y_{il}Y_{it}] &= \lambda_{iul}\mu_{it} + \lambda_{iut}\mu_{il} + \lambda_{ilt}\mu_{iu} - 2\mu_{iu}\mu_{il}\mu_{it}) \\ &= \tilde{\delta}_{iult,N^*} \end{aligned} \tag{2.16}$$

where μ_{it} is computed by (1.8), and λ_{iut}, for example, is computed by (1.17).

Similarly, by using the Lemma 2.2, a general element of the Φ_{i,N^*} matrix may be computed as

$$\begin{aligned} \phi_{iulmt,N^*} &= \mathrm{cov}_M[Y_{iu}Y_{il}, Y_{im}Y_{it}] \\ &= E_M[Y_{iu}Y_{il}Y_{im}Y_{it}] - \lambda_{iul}\lambda_{imt} \end{aligned} \tag{2.17}$$

where, by Lemma 2.2, the fourth order unconditional uncorrected moments have the formula (see also Jowaheer and Sutradhar (2002)) given by

$$\begin{aligned} E_M[Y_{iu}Y_{il}Y_{im}Y_{it}] = {} & \delta_{iul}\sigma_{imt} + \sigma_{ium}\sigma_{ilt} + \sigma_{iut}\sigma_{ilm} \\ & + \tilde{\delta}_{iulm,N^*}\mu_{it} + \tilde{\delta}_{iult,N^*}\mu_{im} \\ & + \tilde{\delta}_{iumt,N^*}\mu_{il} + \tilde{\delta}_{ilmt,N^*}\mu_{iu} \\ & - \sigma_{iul}\mu_{im}\mu_{it} - \sigma_{ium}\mu_{il}\mu_{it} \\ & - \sigma_{iut}\mu_{il}\mu_{im} - \sigma_{ilm}\mu_{iu}\mu_{it} \\ & - \sigma_{ilt}\mu_{iu}\mu_{im} - \sigma_{imt}\mu_{iu}\mu_{il} \\ & + 3\mu_{iu}\mu_{il}\mu_{im}\mu_{it} \end{aligned} \quad (2.18)$$

where $\tilde{\delta}_{iult,N^*}$, for example, is computed by (2.16).

Now if the responses along with covariate information were available from all N individuals of the survey population, then following Sutradhar (2004), one could obtain the finite population (survey population) based consistent estimate for the super population parameter σ_γ^2 by solving the survey population (finite) based GQL estimating equation given by

$$\begin{aligned} & \sum_{i=1}^{N} g^*_{i2,N^*}\left(\sigma_\gamma^2 \middle| \beta, \rho_1, ..., \rho_{T-1}\right) \\ & = \sum_{i=1}^{N} \frac{\partial \eta_i'\left(\sigma_\gamma^2 \middle| \beta, \rho_1, ..., \rho_{T-1}\right)}{\partial \sigma_\gamma^2} \\ & \times \Gamma_{i,N^*}^{-1}\left(u_i - \eta_i\left(\sigma_\gamma^2 \middle| \beta, \rho_1, ..., \rho_{T-1}\right)\right) \\ & = G^*_{2,N^*}\left(\sigma_\gamma^2 \middle| \beta, \rho_1, ..., \rho_{T-1}\right) = 0 \end{aligned} \quad (2.19)$$

where, for known β and ρ_l (l = 1, . . ., T – 1), $\eta_i(\sigma_\gamma^2 | \beta, \rho_1, \ldots, \rho_{T-1}) = E_M[U_i]$ may be computed from (2.12)-(2.13). Let $\sigma^2_{\gamma,N}$ be the finite population based estimate of the super population parameter σ_γ^2, obtained from (2.19).

2.3.2 Finite population based moment estimation of ρ_l (l = 1, . . ., T – 1)

Define $z_{it} = [y_{it} - \mu_{it}]/\sqrt{\sigma_{itt}}$, where μ_{it} and σ_{itt} are computed by (1.8) and (1.9), respectively. Since by (1.16), the model based covariance $\sigma_{iut} = E_M[(Y_{iu} - \mu_{iu})(Y_{it} - \mu_{it})]$ is given by

$$\begin{aligned} \sigma_{iut} = {} & \rho_{|t-u|} W^{-1} \sum_{w=1}^{W}\left[\sqrt{\left[\sigma^*_{iuu}(\gamma_{iw})\sigma^*_{itt}(\gamma_{iw})\right]}\right] \\ & + W^{-1}\sum_{w=1}^{W}\left[\mu^*_{iu}(\gamma_{iw})\mu^*_{it}(\gamma_{iw})\right] - \mu_{iu}\mu_{it} \end{aligned} \quad (2.20)$$

one may then obtain the finite population based moment estimate of ρ_l as

$$\hat{\rho}_l = \frac{\left[\dfrac{NT\sum_{i=1}^{N}\sum_{t=1}^{T-l} z_{it}z_{i,t+l}}{N(T-l)\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}^2} - \sum_{i=1}^{N}\sum_{t=1}^{T-l} \dfrac{\left[W^{-1}\sum_{w=1}^{W}\left\{\mu^*_{it}(\gamma_{iw})\mu^*_{i,t+l}(\gamma_{iw})\right\} - \mu_{it}\mu_{i,t+l}\right]}{N(T-l)\sqrt{\sigma_{itt}\sigma_{i,t+l,t+l}}}\right]}{\left[\sum_{i=1}^{N}\sum_{t=1}^{T-l}\dfrac{W^{-1}\sum_{w=1}^{W}\left[\sqrt{[\sigma^*_{iuu}(\gamma_{iw})\sigma^*_{iit}(\gamma_{iw})]}\right]}{N(T-l)\sqrt{\sigma_{itt}\sigma_{i,t+l,t+l}}}\right]} \quad (2.21)$$

3. SAMPLING DESIGN BASED WEIGHTED GQL ESTIMATING EQUATIONS

3.1 WGQL Estimation of β

It is recognized that in the finite population set up, one does not observe values for all N individuals. Consequently, one can not construct the estimating equation (2.2) to estimate the survey (finite) population parameter β_N. Since in practice, one observes a sample s^* chosen from N individuals based on a suitable complex survey, such as two stage clustered design (D). One then construct this sample based weighted estimating equation to estimate the survey population parameter β_N. As far as the weights are concerned, ω_{is^*} will denote the sampling design weights for the ith unit to be selected in the sample s^*.

We remark here that even if the underlying model is incorrect, this type of weighted estimating equations still produce consistent estimation of the survey population parameters (Pfeffermann 1993).

Now, to obtain a consistent estimate for the survey population parameter β_N under the linear dynamic mixed model, we follow (2.2) and first write the survey estimating function

$$\sum_{i\in s^*} w_{is^*} g_{i1}\left(\beta|\theta,\sigma_\gamma^2,\sigma_\epsilon^2\right)$$

such that

$$E_D\left[\sum_{i\in s^*} w_{is^*} g_{i1}\left(\beta|\theta,\sigma_\gamma^2,\sigma_\epsilon^2\right)\right] = \sum_{i=1}^{N} g_{i1}\left(\beta|\theta,\sigma_\gamma^2,\sigma_\epsilon^2\right) \qquad (3.1)$$

where $E_D(\cdot)$ denote the sampling design based expectation. Consequently, one may obtain consistent estimate for β_N by solving the estimating equation

$$\hat{G}_1\left(\beta|\theta,\sigma_\gamma^2,\sigma_\epsilon^2\right) = \sum_{i\in s^*} w_{is^*} g_{i1}\left(\beta|\theta,\sigma_\gamma^2,\sigma_\epsilon^2\right) = 0 \qquad (3.2)$$

Note that the quantity $g_{i1}(\beta|\,\theta,\ \sigma_\gamma^2,\ \sigma_\epsilon^2)$ in (3.2) is the same GQL function for the i^{th} individual as shown in (2.2). Consequently, the equation in (3.2) is referred to as the weighted (GQL) (WGQL) estimating equation.

As far as the sampling design based estimation for β_N under the binary dynamic mixed model is concerned, we simply replace the GQL function $g_{i1}(\cdot)$ in (3.2) with $g_{i1}^*(\cdot)$ from (2.3). Thus, under the binary dynamic mixed model, one writes the WGQL estimating equation given by

$$\hat{G}_1^*\left(\beta|\rho_1,....,\rho_{T-1},\sigma_\gamma^2\right) = \sum_{i\in s^*} w_{is^*} g_{i1}^*\left(\beta\,|\rho_1,...,\rho_{T-1},\sigma_\gamma^2\right) = 0 \qquad (3.3)$$

Let $\hat{\beta}_{WGQL}$ denote the solution of (3.2) or (3.3). Thus, $\hat{\beta}_{WGQL}$ estimates β_N, the finite population parameter involved in the linear dynamic mixed model, when it is obtained by solving (3.2). Similarly, $\hat{\beta}_{WGQL}$ is an estimate of β_N, the finite population parameter involved in the binary dynamic mixed model, when it is obtained by solving (3.3).

Note that $\hat{\beta}_{WGQL}$, the solution of (3.2) or (3.3), is obtained iteratively by using

$$\hat{\beta}_{WGQL}(m+1) = \hat{\beta}_{WGQL}(m) + \left[\sum_{i\in s^*} w_{is^*} \frac{\partial \mu_i'(\beta)}{\partial\beta} \sum_i^{-1} \frac{\partial \mu_i(\beta)}{\partial\beta'}\right]_m^{-1} \times \left[\sum_{i\in s^*} w_{is^*} \frac{\partial \mu_i'(\beta)}{\partial\beta} \sum_i^{-1} (y_i - \mu_i(\beta))\right]_m \qquad (3.4)$$

where $[]_m$ represents that the quantity in the square bracket is evaluated at $\beta = \hat{\beta}_{WGQL,m}$, $\hat{\beta}_{WGQL,m}$ being the value of $\hat{\beta}_{WGQL}$ at the m^{th} iteration. Furthermore, the covariance matrix of $\hat{\beta}_{WGQL}$ may be estimated as

$$\hat{\text{cov}}\left(\hat{\beta}_{WGQL}\right) = \left[\sum_{i\in s^*} \frac{\partial \mu_i'(\beta)}{\partial\beta} \sum_i^{-1} \frac{\partial \mu_i(\beta)}{\partial\beta'}\right]^{-1} \times \hat{\text{cov}}\left[\sum_{i\in s^*} w_{is^*} \frac{\partial \mu_i'(\beta)}{\partial\beta} \sum_i^{-1} (y_i - \mu_i(\beta)\right] \times \left[\sum_{i\in s^*} w_{is^*} \frac{\partial \mu_i'(\beta)}{\partial\beta} \sum_i^{-1} \frac{\partial \mu_i(\beta)}{\partial\beta'}\right]^{-1} \qquad (3.5)$$

3.2 WGQL Estimation of $\xi = (\theta,\ \sigma_\gamma^2, \sigma_\epsilon^2)'$ for Linear Dynamic Mixed Model

To obtain the sampling design based estimates for the finite population scale parameters in ξ_N, we first write the survey estimating function

$$\sum_{i\in s^*} w_{is^*}\, g_{i2,N^*}\left(\theta,\sigma_\gamma^2,\sigma_\epsilon^2\,\middle|\,\beta\right)$$

such that

$$E_D\left[\sum_{i\in s^*} w_{is^*}\, g_{i2,N^*}\left(\theta,\sigma_\gamma^2,\sigma_\epsilon^2\,\middle|\,\beta\right)\right] = \sum_{i=1} g_{i2,N^*}\left(\theta,\sigma_\gamma^2,\sigma_\epsilon^2\,\middle|\,\beta\right) \qquad (3.6)$$

where

$$\sum_{t=1}^{N} g_{i2,N^*}\left(\theta, \sigma_\gamma^2, \sigma_\in^2 \middle| \beta\right) = \sum_{t=1}^{N} \frac{\partial \sigma_i'}{\partial \xi} \Omega_{iN^*}^{-1} (s_i - \sigma_i)$$

$$= G_{2,N^*}\left(\theta, \sigma_\gamma^2, \sigma_\in^2 \middle| \beta\right)$$

by (2.7). Thus, we obtain the sampling design based WGQL estimator of ξ_N by solving the WGQL estimating equation

$$\hat{G}_{2,N^*}\left(\theta, \sigma_\gamma^2, \sigma_\in^2 \middle| \beta\right) = \sum_{i\in s^*} w_{is^*} g_{i2,N^*}\left(\theta, \sigma_\gamma^2, \sigma_\in^2 \middle| \beta\right)$$

$$= \sum_{i\in s^*} w_{is^*} \frac{\partial \sigma_i'}{\partial \xi} \Omega_{iN^*}' (s_i - \sigma_i)$$

$$= 0 \qquad (3.7)$$

Let $\hat{\xi}_{WGQL}$ denote the solution of (3.7) that estimates the finite population parameter ξ_N.

3.3 WGQL Estimation of $\alpha = (\sigma_\gamma^2, \rho)'$ for Binary Dynamic Mixed Model

3.3.1 WGQL estimation of σ_γ^2

Similarly to Section 3.2, one may obtain the sampling design based WGQL estimate of the finite population parameter $\sigma_{\gamma,N}^2$ under the binary dynamic mixed model, by solving the WGQL estimating equation

$$\hat{G}_{2,N^*}^*\left(\sigma_\gamma^2 \middle| \beta, \rho_1, ..., \rho_{T-1}\right)$$

$$= \sum_{i\in s^*} w_{is^*} g_{i2,N^*}^*\left(\sigma_\gamma^2 \middle| \beta, \rho_1, ..., \rho_{T-1}\right)$$

$$= \sum_{i\in s^*} w_{is^*} \frac{\partial \eta_i'\left(\sigma_\gamma^2 \middle| \beta, \rho_1, ..., \rho_{T-1}\right)}{\partial \sigma_\gamma^2}$$

$$\times \Gamma_{i,N^*}^{-1}\left(u_i - \eta_i\left(\sigma_\gamma^2 \middle| \beta, \rho_1, ..., \rho_{T-1}\right)\right) = 0 \qquad (3.8)$$

where

$$E_D\left[\sum_{i\in s^*} w_{is^*} g_{i2,N^*}^*\left(\sigma_\gamma^2 \middle| \beta, \rho_1 ..., \rho_{T-1}\right)\right]$$

$$= \sum_{i=1}^{N} g_{i2,N^*}^*\left(\sigma_\gamma^2 \middle| \beta, \rho_1 ..., \rho_{T-1}\right) \qquad (3.9)$$

with

$$\sum_{i=1}^{N} g_{i2,N^*}^*\left(\sigma_\gamma^2 \middle| \beta, \rho_1 ..., \rho_{T-1}\right)$$

$$= \sum_{i=1}^{N} \frac{\partial \eta_i'\left(\sigma_\gamma^2 \middle| \beta, \rho_1 ..., \rho_{T-1}\right)}{\partial \sigma_\gamma^2}$$

$$\times \Gamma_{i,N^*}^{-1}\left(u_i - \eta_i\left(\sigma_\gamma^2 \middle| \beta, \rho_1, ..., \rho_{T-1}\right)\right)$$

by (2.19). Let $\hat{\sigma}_{\gamma,WGQL}^2$ denote the solution of (3.8) that estimates the finite population parameter $\sigma_{\gamma,N}^2$, defined by (2.19).

3.3.2 Sampling Design Based Estimation of ρ_l

Note that the finite population based estimate of ρ_l is given by (2.21). This formula is however not applicable as the information from the finite population is not available. To estimate this by using the sample information, we may modify the estimate in (2.21) by using sample observations along with their sampling design weights. This sampling design based moment estimate is given

$$\hat{\rho}_l = \frac{\left[\dfrac{T\sum_{i\in s^*}\sum_{t=1}^{T-l} w_{is^*} z_{it} z_{i,t+l}}{(T-l)\sum_{i\in s^*}\sum_{t=1}^{T} w_{is^*} z_{it}^2} - \sum_{i\in s^*}\sum_{t=1}^{T-l} w_{is^*} \dfrac{\left[W^{-1}\sum_{w=1}^{W}\left\{\mu_{it}^*(\gamma_{iw})\mu_{i,t+l}^*(\gamma_{iw})\right\} - \mu_{it}\mu_{i,t+l}\right]}{(T-l)\sqrt{\sigma_{itt}\sigma_{i,t+l,t+l}}}\right]}{\sum_{i\in s^*}\sum_{t=1}^{T-l} w_{is^*} \dfrac{W^{-1}\sum_{w=1}^{W}\left[\sqrt{[\sigma_{iuu}^*(\gamma_{iw})\sigma_{itt}^*(\gamma_{iw})]}\right]}{(T-l)\sqrt{\sigma_{itt}\sigma_{i,t+l,t+l}}}} \qquad (3.10)$$

4. ANALYSIS OF A LONGITUDINAL BINARY SURVEY DATA: A NUMERICAL ILLUSTRATION

4.1 Statistics Canada SLID Data

Survey of Labour and Income Dynamics (SLID) is a longitudinal household survey conducted by

Statistics Canada from 1993 to 1998, designed to capture changes in the economic well-being of Canadians over time. The target population of this SLID data includes labour force residents of the ten Canadian provinces with the exclusion of the Indian reserves, the military barracks and residential institutions. Since it is impractical to collect employment information from every individuals of this huge target/finite population, Statistics Canada designed a sample based on a stratified cluster sampling scheme. Initially the sample had more than 35,000 individuals. But, as the data were collected longitudinally over a period of six years, the design weights were adjusted every year starting from the second year. We consider the sample s^* with 15,731 individuals who provided employment information for all six years. Their sampling design weights naturally come from the last year of the study. As far as the responses are concerned, let $y_{it} = 1$ denote that the i^{th} individual is unemployed at time t.

Note that the sampling weights are usually denoted by $w_i (i \in s^*)$ or equivalently w_{is^*}, for simplicity. However, the notation for these design weights may further be changed showing the strata and cluster that the individual belongs to. Let

1. L denote the total number of stratum;
2. n_h be the number of clusters in the h^{th} ($h = 1,..., L$) stratum;
3. n_{hc} denote the size of the c^{th} ($c = 1, ..., n_h$) cluster under the h^{th} stratum;
4. w_{hcis^*} represent the design weight for the i^{th} individual to be included in the sample s^* based on his/her origin in the c^{th} cluster of the h^{th} stratum.

Further note that these complex survey information (e.g., L; n_h; and n_{hc}) from the stratified cluster sample, will not be reported in the paper, but they will be available from the author upon request.

Next, with regard to the selection of the covariates, SLID data contains many socioeconomic covariates but for convenience, we consider five important covariates. These covariates are: gender, age, geographic location, education level, and marital status of the individual. While gender, age and geographic location were held as observed in 1993, education level and marital status are considered to be time dependent covariates. As some of these five covariates are categorical with more than two levels, we, for convenience, express these 5 covariates through 12 renamed covariates. Let x_{it} denote the vector for these 12 covariates corresponding to y_{it}. To be specific, the 12 covariates are constructed as follows.

$$x_{it1} = \begin{cases} 0 & \text{for female} \\ 1 & \text{for male} \end{cases}$$

$$(x_{it2}, x_{it3}) \equiv \begin{cases} (0,0) & \text{age group 16 and 24 inclusive} \\ (1,0) & \text{age group 25 and 54 inclusive} \\ (0,1) & \text{age group 55 and 65 inclusive} \end{cases}$$

$(x_{it4}, x_{it5}, x_{it6}, x_{it7})$

$$\equiv \begin{cases} (0,0,0,0) & \text{Atlantic region} \\ (1,0,0,0) & \text{Quebec region} \\ (0,1,0,0) & \text{Ontario region} \\ (0,0,1,0) & \text{Praries} \\ (0,0,0,1) & \text{British Columbia-Alberta region} \end{cases}$$

$$(x_{it8}, x_{it9}) \equiv \begin{cases} (0,0) & \text{low education} \\ (1,0) & \text{medium education} \\ (0,1) & \text{high education} \end{cases}$$

and

$(x_{it10}, x_{it11}, x_{it12})$

$$\equiv \begin{cases} (0,0,0) & \text{married and common law spouse group} \\ (1,0,0) & \text{separated and divorce group} \\ (0,1,0) & \text{widow group} \\ (0,0,1) & \text{single group} \end{cases}$$

4.2 Longitudinal Model and Inferences

Note that $y_{it} = 1$ denote that the ith individual is unemployed at time t. It follows that the repeated binary responses $y_{i1}, ..., y_{it}, ..., y_{iT}$ from the i^{th} individual are likely to be correlated. They may also be affected by the individual random effect $\gamma_i^* \sim \left(0, \sigma_\gamma^2\right)$, but, for simplicity, we assume that individual effects are the same for all 15,731 individuals. Consequently, we need to estimate the effects of the aforementioned 12 covariates on the binary responses $\{y_{it}\}$ after taking the correlations of the repeated data into account. For the purpose, in the absence of random effects, we use

aWGQL estimating equation for β, slightly different from (3.3). This simpler equation has the form

$$\sum_{i\in s^*} w_{is^*} g^*_{i1}\left(\beta\,|\rho_1, ..., \rho_{T-1}\right) = 0 \qquad (4.1)$$

where

$$g^*_{i1}\left(\beta\,|\rho_1, ..., \rho_{T-1}\right) = \frac{\partial \mu_i'}{\partial \beta}\sum\nolimits_i^{-1}(\rho)(y_i - \mu_i)$$

Note that in Section 4.1, the sampling design weights w_{is^*} has been rewritten reflecting a stratified cluster sampling design. Hence we re-write the WGQL estimating equation (4.1) as

$$\sum_{i\in s^*} w_{is^*} z^*_{is^*} \equiv \sum_{i=1}^{n_{hc}} w_{hcis^*} z^*_{hcis^*} = z^*_{hc} = 0 \qquad (4.2)$$

where

$$z^*_{is^*} = \frac{\partial \mu_i'}{\partial \beta}\sum\nolimits_i^{-1}(\rho)(y_i - \mu_i)$$

for $i \in s^*$. Let $\hat{\beta}_{WGQL(F)}$ (WGQL regression estimate under finite population model for longitudinal data) be the solution of (4.2) which may be obtained iteratively by using

$$\hat{\beta}_{WGQL(F)}(m+1) = \hat{\beta}_{WGQL(F)}(m) + \left[\sum_{i=1}^{n_{hc}} w_{hcis^*} \frac{\partial \mu_i'}{\partial \beta}\sum\nolimits_i^{-1}(\rho)\frac{\partial \mu_i}{\partial \beta'}\right]_m^{-1} \times \left[\sum_{i=1}^{n_{hc}} w_{hcis^*} \frac{\partial \mu_i'}{\partial \beta}\sum\nolimits_i^{-1}(\rho)(y_i - \mu_i)\right]_m \qquad (4.3)$$

where ρ represents all ρ_l (l = 1, …, T – 1). In the absence of random effects, by following (3.10), these lag correlations may now be estimated by

$$\hat{\rho}_l = \frac{\dfrac{\sum_{i=1}^{n_{hc}}\sum_{t=1}^{T-l} w_{hics^*} z_{it} z_{i,t+l}}{(T-l)\sum_{i=1}^{n_{hc}} w_{hcis^*}}}{\sum_{i=1}^{n_{hc}}\sum_{t=1}^{T} w_{hcis^*} z^2_{it} \Big/ T\sum_{i=1}^{n_{hc}} w_{hcis^*}} \qquad (4.4)$$

where the weights are used in new notation w_{hcis^*}.

As far as the estimation of the standard errors of the estimates of the components of β is concerned, this may be done by estimating the covariance matrix of $\hat{\beta}_{WGQL(F)}$ as

$$\hat{\text{cov}}\left(\hat{\beta}_{WGQL(F)}\right) = \left[\sum_{i=1}^{n_{hc}} w_{hcis^*} \frac{\partial \mu_i'}{\partial \beta}\sum\nolimits_i^{-1}(\rho)\frac{\partial \mu_i}{\partial \beta}\right]^{-1} \times \hat{\text{cov}}\left[\sum_{i=1}^{n_{hc}} w_{hcis^*} \frac{\partial \mu_i'}{\partial \beta}\sum\nolimits_i^{-1}(\rho)\,(y_i - \mu_i)\right] \times \left[\sum_{i=1}^{n_{hc}} w_{hcis^*} \frac{\partial \mu_i'}{\partial \beta}\sum\nolimits_i^{-1}(\rho)\,\frac{\partial \mu_i}{\partial \beta}\right]^{-1} \qquad (4.5)$$

where, the middle term in (4.5), i.e.,

$\hat{\text{cov}}\left(\sum_{i=1}^{n_{hc}} w_{hcis^*} z^*_{hcis^*}\right)$, may be computed as

$$\hat{\text{cov}}\left(\sum_{i=1}^{n_{hc}} w_{hcis^*} z^*_{hcis^*}\right) = \left[\sum_{h=1}^{L}\left\{n_h(n_h-1)^{-1}\sum_{c=1}^{n_h}\left(z^*_{hc} - \bar{z}^*_h\right)\left(z^*_{hc} - \bar{z}^*_h\right)'\right\}\right] \qquad (4.6)$$

where $\bar{z}^*_h = \sum_{c=1}^{n_h} z^*_{hc} \Big/ n_h$ is the $p \times 1$ mean vector.

The regression estimates and their standard errors obtained from (4.3) and (4.5)-(4.6), under the finite population model, are reported in the last two columns (4 and 5) of Table 1. In columns 2 and 3 of the same table, we have also reported the regression estimates and their standard errors under the infinite population set up, that is, by computing the estimates and their standard errors under the assumption that the 15,731 individuals were selected from an infinite population based on the simple random sampling design so that w_{is^*} is the same for all individuals. The lag correlations obtained by (4.4) under the present finite population model are reported in column 4.

Table 1. Estimates of regression and their estimated standard errors, as well as estimates of autocorrelations under both infinite and finite population set up for complete SLID data.

	Set up			
	Infinite population		Finite population	
Parameters	Estimate	SE	Estimate	SE
Male vs Female (x_1)	–0.638	0.066	–0.502	0.074
Age group 2 vs 1 (x_2)	–1.688	0.057	–1.285	0.087
Age group 3 vs 1 (x_3)	–2.489	0.115	–1.864	0.154
Quebec vs Atlantic (x_4)	–0.762	0.078	–1.249	0.157
Ontario vs Atlantic (x_5)	–1.052	0.088	–1.528	0.110
Praries vs Atlantic (x_6)	–1.702	0.114	–2.061	0.136
BC & Alberta vs Atlantic (x_7)	–1.482	0.169	–1.955	0.191
Education medium vs low (x_8)	–1.681	0.058	–1.589	0.092
Education high vs low (x_9)	–2.446	0.153	–2.609	0.236
Marital status 2 vs 1 (x_{10})	0.193	0.096	0.243	0.142
Marital status 3 vs 1 (x_{11})	–0.688	0.248	–0.480	0.379
Marital status 4 vs 1 (x_{12})	–0.566	0.077	–0.343	0.146
ρ_1	0.393	–	0.360	–
ρ_2	0.243	–	0.202	–
ρ_3	0.155	–	0.123	–
ρ_4	0.129	–	0.089	–
ρ_5	0.126	–	0.087	–

4.3 Estimates and their Interpretation

The lag 1 correlation is found to be 0.360 which is not too small and hence can not be ignored in efficient estimation of β. The other lag correlations show a decaying pattern as expected.

We now interpret the WGQL estimates of the components of the β vector. The negative value –0.502 for the gender effect indicates that the male has lower probability of an all-year unemployment as compared to the female. The negative values –1.285 and –1.864 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 the Praries had smallest probability of an all-year unemployment during 1993 to 1998 followed by BC and Alberta, Ontario, Quebec and Atlantic provinces. This follows from the fact that the regression estimates for Quebec, Ontario, BC and Alberta, and Prairies are found to be –0.1.249, –1.528, –1.955, and –2.061 respectively. The larger negative value –2.609 for β_9 as compared to β_8 = –1.589 indicates that as the education level gets higher, the probability of an all-year unemployment gets smaller. Finally, with regard to the marital status, the positive value 0.243 for β_{10} means that the separated and divorced individuals have higher 24 probability of all-year unemployment as compared to the married and common law spouse group. Similarly, the widowed had less probability of an all-year unemployment as compared to the single but never married individual.

Note that the simple random sampling (SRS) design based regression estimates reported in

column 2, are in general different than the stratified clustered sampling (SCS) based estimates (in column 4) but their patterns are similar. As expected, the SCS design based estimates have larger standard errors as compared to the SRS design based estimates, but the former estimates are known to be less biased than the SRS design based estimates.

ACKNOWLEDGEMENTS

The authors are grateful to Bhagavan Sri Sathya Sai Baba for providing opportunities to carry out a part of this research at the Sri Sathya Sai University. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Amemiya, T. (1985). *Advanced Econometrics.* Harvard University Press, Cambridge.

Binder, D.A. (1993). On the variance of asymptotically normal estimators from complex surveys. *Int. Statist. Rev.,* **51**, 279-292.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Rev. Econ. Statist.*, **58**, 277-298.

Bun, M.J.G. and Carree, M.A. (2005). Bias correction estimation in dynamic panel data models. *J. Busi. Eco. Statist.*, **23**, 200-210.

Chamberlain, G. (1992). Sequential moment restrictions in panel data-comment. *J. Busi. Eco. Statist.,* **10**, 20-26.

Godambe, V.P. and Thompson, M.E. (1986). Parameters of super-population and survey population: Their relationships and estimation. *Int. Statist. Rev.,* **54**, 127-138.

Hsiao, C. (2003). *Analysis of Panel Data.* Cambridge University Press, Cambridge, U.K.

Imbens, G.W. (2002). Generalized method of moments and empirical likelihood. *J. Busi. Eco. Statist.*, **20**, 493-506.

Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **93**, 720-729.

Pfeffermann, D. (1993). The role of sampling weights when modelling survey data. *Int. Statist. Rev.,* **61**, 317-337.

Qaqish, B.F. (2003). A family of multivariate binary distribution for simulating correlated binary variables with specified marginal means and correlations. *Biometrika,* **90**, 455-463.

Sutradhar, B.C. (2003). An review on regression models for discrete longitudinal responses. *Statist. Sci.,* **18**, 377-393.

Sutradhar, B.C. (2004). On exact quasilikelihood inference in generalized linear mixed models. *Sankhya: The Indian Journal of Statistics,* **66**, 261-289.

Sutradhar, B.C. and Kovacevic, M. (2000). Analyzing ordinal longitudinal survey data: Generalized estimating equations approach. *Biometrika,* **87**, 837-848.

Sutradhar, B.C., Rao, R.P. and Pandit, V.N. (2008). Generalized method of moments versus generalized quasilikelihood inferences in binary panel data models. *Sankhya: The Indian Journal of Statistics,* **B70**, 34-62.