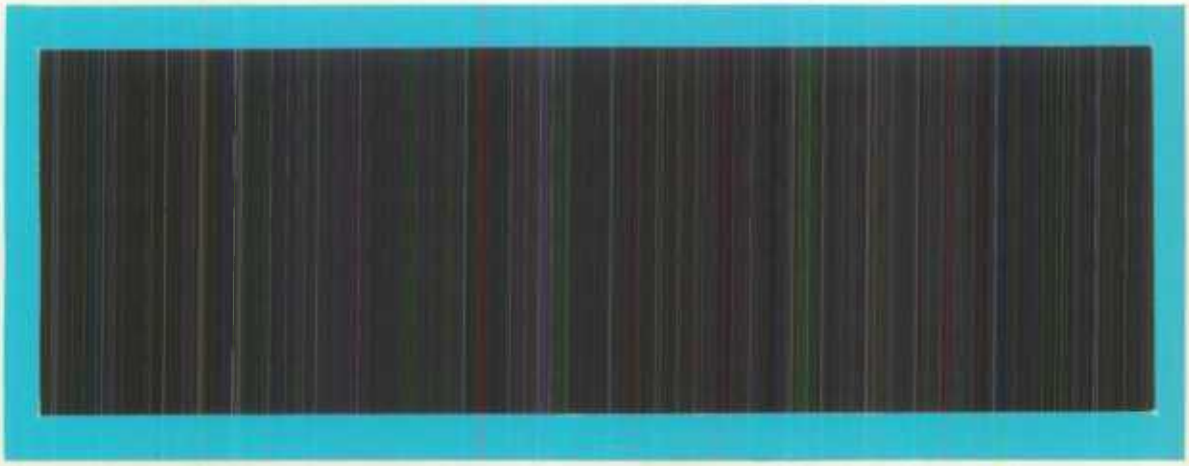
Statistics Canada Statistique Canada



Methodology Branch

Social Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes sociales

Canada

WORKING PAPER

METHODOLOGY BRANCH

ANALYSIS OF MULTIVARIATE ORDINAL DATA FOR LONGITUDINAL SURVEYS: PART I

SSMD - 99-005 E

Brajendra C. Sutradhar

Memorial University of Newfoundland

A Report Prepared for Statistics Canada

Revised: February 25, 1999

Abstract

Longitudinal survey data may comprise of ordinal polytomous repeated observations and a set of multi-dimensional covariates, for a large number of individuals. One of the main goals of the longitudinal survey is to see what happens to individuals or households over time. More precisely, one may like to describe the marginal expectation of the ordinal polytomous outcome variable as a function of the covariates while accounting for the structural (cross-sectional) as well as longitudinal correlations. The structural correlations arise because of the polytomous nature of the response variable, and the longitudinal correlations arise because of the repeatation of the structurally correlated responses over time. In this report, we develop a robust longitudinal correlations structure based generalized estimating equations approach to deal with multivariate polytomous survey data. This we do to analyze longitudinal survey data, such as SLID data collected by Statistics Canada. Details are given for the construction of the mean vector and structural and longitudinal correlations that are used in the development of the estimating equations. The regression estimates, that is, the estimates of the covariate effects, are shown to be consistent for the corresponding regression parameters.

Résumé

Les données d'enquêtes longitudinales peuvent comprendre des observations répétées polytomiques ordinales et un ensemble de covariables multidimensionnelles pour un grand nombre d'individus. L'un des principaux objectifs de l'enquête longitudinale est de voir ce qui arrive aux individus ou aux ménages avec le temps. Plus précisément, une telle analyse peut décrire l'espérance mathématique marginale de la variable de résultat polytomique ordinale en tant que fonction des covariables, tout en représentant les corrélations structurelles (transversales) aussi bien que longitudinales. Les corrélations structurelles apparaissent en raison de la nature polytomique de la variable de réponse, alors que les corrélations longitudinales surviennent en raison de la répétition des réponses structurellement corrélées avec le temps. Nous développons ici une approche généralisée des équations d'estimation basée sur une structure robuste de corrélations longitudinales pour traiter les données d'enquêtes polytomiques multivariables. Cela nous permet d'analyser des données d'enquêtes longitudinales comme les données de l'EDTR recueillies par Statistique Canada. Sont présentés des détails sur la construction du vecteur des moyennes et les corrélations structurelles et longitudinales qui sont utilisées dans le développement des équations d'estimation. Les estimations de régression, c'est-à-dire les estimations des effets des covariables, s'avérent cohérentes pour les paramètres de régression correspondants.

CONTENTS

1 Introduction

Over the last few years, Statistics Canada has been conducting a number of large-scale longitudinal surveys including the Survey of Labour and Income Dynamics (SLID), the National Population Health Survey (NPHS), and the National Longitudinal Survey of Children and Youth (NLSCY). In general, these longitudinal data are comprised of multi-dimensional repeated observations of a vector outcome and a set of multi-dimensional covariates under each of many independent households or individuals. One of the main goals of the longitudinal survey is to see what happens to households and families or individuals over time. More precisely, one of the main objectives is to describe the marginal expectation of the outcome variable as a function of the covariates while accounting for the structural as well as longitudinal correlations. In the multivariate set-up, there are two types of structural correlations. First, at a given point of time, the multivariate responses for an individual are correlated; and second, the responses of the individuals in a household or family may be correlated. Next, the longitudinal correlations arise because of the repeatation of the structurally correlated responses over a period of time. But, as there is no unique way to model such structural and longitudinal correlations, the regression analysis becomes extremely complicated. Further problems may be mounted because of the nature of the complex design used to collect such multi-dimensional longitudinal responses, mainly under the situations when the composition of a household or family changes over time.

In this report, we develop multivariate regression approaches for two types of longitudinal survey data. In the first case, it will be assumed that a cluster or household is selected as the unit of interest based on a suitable survey design. The data collected from the individuals of the cluster at different points of time will be considered as a single multi-dimensional response. For simplicity, it will, however, be assumed that the composition of the household remains the same over the period of time, although the household sizes may be different. In the second case, the individual will be treated as the unit of interest from the longitudinal point of view, and the data collected from the individual over a period of time will be considered as a single multi-dimensional response. Further, it

will be assumed that the longitudinal survey data is complete. The problems of variable households composition over the period of time, as well as the problems of missing data, if any, will be dealt with in the next report.

The specific plan of the report is as follows. An overview of the recent longitudinal surveys conducted by Statistics Canada is given in Section 2. The weighting issues for the longitudinal surveys is described in Section 3, necessarily in a brief, overview format. In order to analyze the longitudinal survey data, Section 4 deals with the formulation of the problems in a regression set-up, in the context of SLID data, for example. Survey weights based estimation steps are provided in the same section. The computational formula for the standard error of the estimator is also given in Section 4. In Section 5, we provide the rationale for the estimating equations that we have used in Section 4. Details are given for the construction of the mean vector and structural and longitudinal correlations that are used in the development of the estimating equations. We conclude the report in Section 6.

2 Longitudinal Surveys by Statistics Canada

A few years ago, Statistics Canada has begun several large-scale longitudinal surveys, such as SLID in 1993/1994, NPHS in 1994/1995, and NLSCY also in 1994/1995. The basic objectives of these longitudinal surveys is to see what happens to households and families or individuals over time in different contexts.

More specifically, the Survey of Labour and Income Dynamics (SLID), for example, is a longitudinal survey of households or individuals designed by Statistics Canada, to measure the changes that take place in the level of socio-economic well-being of the individuals. The sample for this survey was selected in 1993, which is divided into two ovelapping panels that remain in place for a period of six years each. The collection of the first wave of data (i.e., from the first panel) began in 1994 and the second wave was introduced in 1997. Each panel consists of 15000 households (approximately

40000 individuals). A new panel will be subsequently selected every three years to replace the older of the two panels. Every year, information is collected on the panel members' labour market activity and income during the preceeding year. In this problem, it may be of interest to determine (1) the causes of movement between unemployment and employment by looking at the spells of unemployment and characteristics that one may relate to the length of those spells; (2) what are the measured variables that explain marriage duration; and (3) the distribution of lengths of welfare spells and to determine what factors do affect such spells. Some common characteristics those may be related to the unemployment spells, the marriage duration, and the welfare spells are age, sex, income, and the education level. The main purpose of such longitudinal study is to examine the effect of the characteristics or covariates on the responses, namely, on the unemployment spells, the marriage duration, and the welfare spells.

Another large-scale longitudinal survey undertaken by Statistics Canada is the National Population Health Survey (NPHS). This survey is designed to collect data from a longitudinal sample of respondents about their health status, the use of health services and medications, and their life style as well as their demographic and economic information. The results from this survey will help to understand, among other things, the relationship between health status and health care utilization, including alternative as well as traditional services. The first 12-month cycle of data collection began in 1994 from a sample of about 26000 households. From each household, one person aged twelve years and over was selected for an in depth study and became part of the longitudinal panel. It was decided that the data will be collected from this panel every two years for two decades. Thus, Statistics Canada currently has two waves of data under this NPHS.

As far as the National Longitudinal Survey of Children and Youth (NLSCY) is concerned, Statistics Canada has already conducted two waves of survey in 1994-95 and 1996-97. The sample consists of 23000 children. Their age range from newborn to eleven years old. The survey will be repeated at two-year intervals to follow these children as they grow to reach adulthood. This survey covers a braod range of characteristics and factors affecting the growth of children and development

For an overall idea about the nature of the SLID, NPHS and NLSCY and other longitudinal survey data, we refer to Statistics Canada reports, for example, prepared by Lawless (1997), Latouche and Michaud (1995), Hapuarachci (1996), and Tambay and Catlin (1995). Latouche and Michaud (1995), in particular, discuss different steps involved in collecting SLID data, which is helpful in developing both cross-sectional and longitudinal survey weights. Similarly, Tambay and Catlin (1995) discuss the data collection steps for the National Population Health Survey (NPHS). But, there does not appear any adequate discussion to analyse such longitudinal survey data. This report is one step toward the methodological developments for analyzing the longitudinal survey data collected by Statistics Canada. More specifically, in this report, we develop a multivariate polytomous ordinal regression approach to analyze the longitudinal survey data, for example the SLID data.

3 Weighting Issues for the Longitudinal Surveys

The SLID, for example, follow individuals and households, tracking their labour market activities and changes in income and family circumstances. To begin with, SLID sample was a subsample of the Canadian Labour Force Survey (LFS). The LFS uses a multi-stage stratified sample design based on an area frame with dwellings as ultimate sampling units. SLID actually follows individuals through time, but household characteristics are also of interest. Consequently, the use of a complex survey design combined with cross-sectional expectations complicate the different steps in the weighting process. Lavallee and Hunter (1992) have addressed the problem of making the SLID longitudinal sample representative for cross-sectional estimation. These authors have discussed the determination of the basic weights for the SLID sample, as well as their nonresponse and post-stratification adjustments. The problem of determining the basic weights, for the purpose of cross-sectional estimation, is complicated by the fact that cohabitants and new entrants can be part of the sample at any wave of interviewing by joining a longitudinal household.

In the present report, we, however, concentrate mainly to the longitudinal estimation. Thus, we consider those individuals who are assumed to be in the sample for the complete duration of the survey, with some possiblities that some of the individuals may be missing from the survey occassionally or for ever. This later problem of missing responses will be dealt in the next report. For simplicity, we now assume that the survey weights for the longitudinal individuals are known, for example, from Lavallee and Hunter. Let w_{is^*} denote the survey weight for the ith $(i = 1, \ldots, I)$ individual, which may depend on the sample s^*, say, from an appropriate finite population. In the present approach, the information collected from an individual over the period of time will be considered as a single piece of multi-dimensional information.

In some situations, it may be convenient to deal with the longitudinal households as the units of interest. In such cases, the information obtained from the individuals of the household over a period of time will be considered as a single piece of multi-dimensional information. Here, in general, the number of individuals may vary from household to household. Let w_{hs^*} be the survey weight for the hth $(h = 1, \ldots, H)$ household, which may depend on the sample s^*.

We note here that the longitudinal survey weights are usually choosen in such way that the sampling design provides consistent and asymptotically normal estimators of certain population totals, and associated standard errors.

4 Analyzing Longitudinal Survey Data

We now proceed to develop a regression methodology to analyze the longitudinal survey data described in the previous secions. For convenience, we discuss the methodology in the context of SLID data. Note that each of the responses of this SLID data can be categorized into more than two ordinal groups. For example, the unemployment spells can be divided into ordinal groups such as 0-3 months, 3-6 months, 6-9 months, ..., and so on. The individual response at a given year will fall into any of these ordinal groups. As the data are collected longitudinally, the responses of the same

individual for another year may fall into the same or any other group. This clearly demonstrates the need for the development of multivariate regression methods for polytomous ordinal data, which we describe below.

4.1 A Multivariate Regression Approach

Suppose that, attached to all units of a finite population of size I, we have measurements (x_i, y_i^*) made on a matrix of covariates, X, and a response vector, Y^*. More specifically, at a given year t, let $Y_{it}^* = [Y_{it1}^{*T}, \ldots, Y_{itr}^{*T}, \ldots, Y_{its}^{*T}]^T$ represent the response vector for the ith individual, where Y_{itr}^* is a $(J_r - 1)$-dimensional polytomous response vector for the ith individual, corresponding to the rth $(r = 1, \ldots, s)$ variable. For SLID data, consider 'unemployment spell' as the first response variable $(r = 1)$. Now, if the response of the 50th individual, for example, at year $t = 2$, is considered to fall into any of the 10 ordinal groups (say), then $Y_{50,2,1}^*$ is the $J_r - 1 = 9$ dimensional response vector containing one 1 and 8 zeros, and so on.

Next, suppose that $Y_i^* = (Y_{i1}^{*T}, \ldots, Y_{it}^{*T}, \ldots, Y_{iT_0}^{*T})^T$ is the combined response vector for the ith individual collected from T_0 number of years. Further, let X_{itr} denote a p-dimensional possibly time dependent marginal covariate corresponding to the tth $(t = 1, \ldots, T_0)$ time and rth $(r = 1, \ldots, s)$ variable for the ith $(i = 1, \ldots, I)$ individual. Let β_r be the covariate effect of X_{itr} on the $(J_r - 1)$-dimensional response vector Y_{itr}^* and $\beta = (\beta_1^T, \ldots, \beta_r^T, \ldots, \beta_s^T)^T$ denote the $p\sum_{r=1}^{s}(J_r - 1)$-dimensional vector of all regression parameters.

Let $\mu_i^* = (\mu_{i1}^{*T}, \ldots, \mu_{it}^{*T}, \ldots, \mu_{iT_0}^{*T})^T$ denote the expectation vector of Y_i^*. Also let $\Sigma_{w\cdot i}^*$ denote a working covariance matrix that represents both of the structural (due to the multi-dimensional nature of the response) as well as the longitudinal (due to the repeatation of the response over time) correlations for the ith individual. The purpose of the proposed regression methodology is to estimate the regression effects β after taking the sampling design (discussed in Section 3), as well as the above structural and longitudinal correlations of the responses, into account. The construction of the $\mu_i^*(\beta)$ vector and the $\Sigma_{w\cdot i}^*$ matrix is discussed in details in Section 5.

4.2 Estimating Equations

Note that in the present set-up, we do not observe values for all the population units but only for those in a sample drawn from the finite population according to some well-defined sampling scheme. We are interested in estimating β and testing certain hypotheses about β.

Suppose that, if we had values for the whole finite population, we could obtain a consistent estimator of β by solving the estimating equations

$$S^*(\beta) = \sum_{i=1}^{I} u_i^*(\beta) = 0 \tag{4.1}$$

where $u_i^*(\beta)$ has kth ($k = 1, \ldots, p$, say) component u_{ik}^*, say. Further suppose that the sample design provides consistent, asymptotically normal estimators of population totals, and associated standard errors. Then, since $S^*(\beta)$ is a vector of population totals for fixed β, similar to Rao, Scott and Skinner (1997), we can produce an estimator of $S^*(\beta)$ as

$$\hat{S}^*(\beta) = \sum_{i \in s^*} w_{is^*} u_i^*(\beta), \tag{4.2}$$

where the survey weights, w_{is^*}, may depend on the sample s^*. This approach was suggested by Binder (1983) for generalized linear models and any survey design. Note that although it is not essential, it is helpful to have some ideas about u_{ik}^* in the finite population level. We show in Section 5.1.2 that under the special covariance structure $\Sigma_{w_i}^*$, the $u_{ik}^*(\beta)$ vector in (4.1) or (4.2) may be expressed as

$$u_i^*(\beta) = W_i^{*T} \Sigma_{w_i}^{*-1} (Y_i^* - \mu_i^*(\beta)), \tag{4.3}$$

with $W_i^{*T} = \partial (Y_i^* - \mu_i^*(\beta))^T / \partial \beta$, showing $u_{ik}^*(\beta)$ as the kth component of the ith individual, at finite population level.

Finally, the sample estimator, $\hat{\beta}$, is obtained by solving $\hat{S}^*(\hat{\beta}) = 0$, where

$$\begin{aligned} \hat{S}^*(\beta) &= \sum_{i \in s^*} w_{is^*} u_i^*(\beta) \\ &= \sum_{i \in s^*} w_{is^*} W_i^{*T} \Sigma_{w_i}^{*-1} (y_i^* - \mu_i^*(\beta)). \end{aligned} \tag{4.4}$$

Note that it is customary to obtain $\hat{\beta}$ from $\hat{S}^*(\hat{\beta}) = 0$ by using the well-known Newton Raphson iteration method. Given the value $\hat{\beta}(m)$ at the mth iteration, $\hat{\beta}(m+1)$ is obtained as

$$\hat{\beta}(m+1) = \hat{\beta}(m) + [F(\beta)]_m^{-1} \left[\sum_{i \epsilon s^*} w_{is^*} W_i^{*T} \Sigma_{w_i}^{*-1} (Y_i^* - \mu_i^*(\beta)) \right]_m , \tag{4.5}$$

where

$$\begin{aligned} F(\beta) = -\frac{\partial \hat{S}^*}{\partial \beta'} &= \sum_{i \epsilon s^*} w_{is^*} \frac{\partial u_i^*(\beta)}{\partial \beta^T} \\ &= \sum_{i \epsilon s^*} w_{is^*} W_i^{*T} \Sigma_{w_i}^{*-1} W_i^*, \end{aligned} \tag{4.6}$$

with $W_i^* = \partial \mu_i^{*T}(\beta)/\partial \beta$ as in (4.3), and $[.]_m$ denotes that the expression within the brackets is evaluated at $\hat{\beta}(m)$.

4.3 Asymptotic Properties of $\hat{\beta}$

Under suitable conditions [cf. Binder (1983) for details, see also Rao, Scott and Skinner (1997)], $\hat{\beta}$ is asymptotically normal with mean β, and cov$(\hat{\beta})$ can be consistently estimated by

$$\hat{V}(\hat{\beta}) = [F(\hat{\beta})]^{-1} \hat{V}_{s^*}(\hat{\beta}) [F(\hat{\beta})]^{-1}, \tag{4.7}$$

where $F(\beta)$ is given by (4.6), and $\hat{V}_{s^*}(\hat{\beta})$ is the estimated covariance of $\hat{S}^*(\beta)$ under the specified survey design evaluated at $\beta = \hat{\beta}$. Note that $\hat{V}_{s^*}(\hat{\beta})$ may be obtained from the standard survey variance estimator for a total since $\hat{S}^*(\beta)$, given in (4.2), is the estimator of the total $S^*(\beta)$ given by (4.1).

5 Details on Methodological Development

In this section, we describe in details how to compute the vector and matrix components necessary to contruct the estimating equations (4.2), for example. We, however, first, discuss the rationale for the use of such estimating equations in the context of longitudinal survey data.

5.1 Estimating Equations

In this section, we discuss the rationale for the estimating equations for two types of sampling units from longitudinal point of view. Although, in Section 4, we have introduced the estimating equations for β for the case when individuals are units of interest, we, however, first, consider a general case where a household is the unit of interest from longitudinal point of view. In this case, information obtained from the individuals of the household over a period of time is considered to be a single multi-dimensional information. Second, we consider individuals themselves as the units of interest from longitudinal point of view. In this approach, the information collected from an individual over the period of time will be considered as a single piece of multi-dimensional information. The second case was discussed in Section 4, which may be obtained from the first case by using the single membered household in place of the households with variable sizes.

5.1.1 Cluster or household as the unit of interest

Suppose that, attached to all units of a finite population of size H, we have measurements (x_h, y_h) made on a matrix of covariates, X, and a response vector, Y. We assume that for a given value of X, Y is generated by a random process described in Section 5.2 and 5.3, with mean vector

$$E(Y_h) = \mu_h = \mu(X_h, \beta) \tag{5.1}$$

and suppose that we have in mind some working model for the covariance matrix, say

$$\text{var}(Y_h) = \Sigma_{wh} = \Sigma_w(\mu_h) \tag{5.2}$$

for $h = 1, \ldots, H$.

Now by similar arguments as in Section 4.2, we could obtain a consistent estimator of β by solving the estimating equations

$$S(\beta) = \sum_{h=1}^{H} u_h(\beta) = 0, \tag{5.3}$$

provided u_{hk}, the kth component of $u_h(\beta)$, is a distinct component for the hth household at the finite population level. Further, since $S(\beta)$ is a vector of population totals for fixed β, similar to (4.2), we can produce an estimator of $S(\beta)$ as

$$\hat{S}(\beta) = \sum_{h \epsilon s^*} w_{hs^*} u_h(\beta), \tag{5.4}$$

where the survey weights, w_{hs^*}, may depend on the sample s^*. We now explore the nature of the components of $u_h(\beta)$ at the finite population level. In the next section, we will similarly explore the in depth nature of $u^*_{ik}(\beta)$ at the finite population level. This $u^*_{ik}(\beta)$ was used in (4.3) to construct the estimating equations for β, for the case when individuals are considered to be the units of interest.

Suppose that for $m = 1, \ldots, n_h$, ρ_{hm} is the $\sum_{r=1}^{s}(J_r - 1) \times \sum_{r=1}^{s}(J_r - 1)$ structural correlation matrix for the mth individual of the hth household. Here we have assumed that the rth ($r = 1, \ldots, s$) ordinal variable has J_r categories. The modelling for the ρ_{hm} matrix is discussed in Section 5.3. We further assume that ϕ is the cross correlation between any two individuals in a given household. This structural correlation is usually referred to as the familial correlation. Next, suppose that α denotes the correlation between any two values collected at two different time points for the mth individual of the hth household. If the data is collected for T_0 times for a household, then the $n_h T_0 \sum_{r=1}^{s}(J_r - 1) \times n_h T_0 \sum_{r-1}^{s}(J_r - 1)$ working correlation matrix for the Y_h vector defined in (5.1), can be written as

$$\begin{aligned} C_{wh} &= \text{corr}(Y_h) \\ &= \begin{bmatrix} D_{h1} & 0 & \cdots & 0 \\ 0 & D_{h2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & D_{hT_0} \end{bmatrix} + \alpha 1_d 1_d^T, \end{aligned} \tag{5.5}$$

where $d = n_h T_0 \sum_{r=1}^{s}(J_r - 1)$, 1_d is the d-dimensional unit vector, and for all $t = 1, \ldots, T_0$, D_{ht} is the

$n_h \sum_{r=1}^{s}(J_r - 1) \times n_h \sum_{r=1}^{s}(J_r - 1)$ matrix given by

$$D_{ht} = \begin{bmatrix} \rho_{h1} & \phi 1_{d_1} 1_{d_1}^T & \cdots & \phi 1_{d_1} 1_{d_1}^T \\ & \rho_{h2} & \cdots & \phi 1_{d_1} 1_{d_1}^T \\ & & \ddots & \vdots \\ & & & \rho_{hn_h} \end{bmatrix} - \alpha 1_{d_2} 1_{d_2}^T, \tag{5.6}$$

with $d_1 = \sum_{r=1}^{s}(J_r - 1), d_2 = n_h \sum_{r=1}^{s}(J_r - 1) = n_h d_1$. In (5.5) $Y_h = [Y_{h_1}^T, \ldots; Y_{ht}^T, \ldots, Y_{hT_0}^T]^T$ with $Y_{ht} = [Y_{ht1}^T, \ldots, Y_{htm}^T, \ldots, Y_{htn_h}^T]^T$, where $Y_{htm} = [Y_{htm_1}^T, \ldots, Y_{htmr}^T, \ldots, Y_{htms}^T]$, with Y_{htmr} as the $(J_r - 1)$-dimensional polytomous responses for the rth $(r = 1, \ldots, s)$ variable. It now follows from (5.5) that

$$C_{w\,h}^{-1} = D_h^{-1} - \{\alpha/\eta(\alpha)\} D_h^{-1} 1_d 1_d^T D_h^{-1},$$

where $\eta(\alpha) = \{1 + \alpha 1_d^T D_h 1_d\}$, and where $D_h^{-1} = \oplus D_{ht}^{-1}$ $(t = 1, \ldots, T_0)$.

Next write,

$$\begin{aligned} s_{ht} &= y_{ht} - \mu_{ht}(\beta) \\ s_h &= [s_{h1}^T, \ldots, s_{ht}^T, \ldots, s_{ht_0}^T]^T, \\ \Sigma_{w\,h}^{-1} &= A_h^{-\frac{1}{2}} C_{w\,h}^{-1} A_h^{-\frac{1}{2}}, \end{aligned}$$

where $A_h = \oplus A_{ht}$ with $A_{ht} = \mathrm{cov}(Y_{ht})$. Also write

$$\begin{aligned} B_{ht} &= A_{ht}^{-\frac{1}{2}} D_{ht}^{-1}, \\ W_h^T &= \left(\frac{\partial s_h^T}{\partial \beta}\right) \\ &= (W_{h1}^T, \ldots, W_{ht}^T, \ldots, W_{hT_0}^T), \end{aligned}$$

where W_{ht}^T is the $p \times n_h \sum_{r=1}^{s}(J_r - 1)$ matrix. Now using $u_h(\beta) = W_h^T \Sigma_{w\,h}^{-1} s_h$, the estimating equation

(5.3) may be written as

$$\begin{aligned} S(\beta) &= \sum_{h=1}^{H} u_h(\beta) \\ &= \sum_{h=1}^{H} W_h^T \Sigma_{wh}^{-1} s_h \\ &= \sum_{h=1}^{H} \sum_{t=1}^{T_0} W_{ht}^T \tilde{s}_{ht} = 0, \end{aligned} \tag{5.7}$$

where $\tilde{s}_{ht} = B_{ht}^{-1} \left(s_{ht} - \frac{\alpha}{\eta(\alpha)} \sum_{v=1}^{T_0} B_{hv}^{-1} s_{hv} \right)$, yielding

$$u_{hk}(\beta) = \sum_{t=1}^{T_0} W_{htk}^T \tilde{s}_{ht}, \tag{5.8}$$

where W_{htk}^T is the kth ($k = 1, \ldots, p$) row of the $p \times n_h \sum_{r=1}^{s} (J_r - 1)$ matrix, W_{ht}^T.

The above computations show that in order to use the design weights for the inference about β, it is essential to assume that for given β at the finite population level, there exists p-components of a vector $u_h(\beta)$, where the kth component is defined by (5.8). One then includes the $u_h(\beta)$ vector in the sample s^* with weight w_{hs^*} as in (5.4).

5.1.2 Individual as the unit of interest

In many Statistics Canada longitudinal survey data, in particular in the SLID data, the individuals are considered as the units of interest from the longitudinal point of view. This is because due to its dynamic nature over time, the household is difficult to use as a tool for longitudinal analysis, even if it corresponds quite closely to the sampling unit [cf. Latouche and Michaud (1995)]. Moreoever, the individuals in the longitudinal sample help create new households by leaving or welcoming new members into their original households. Consequently, it is much better to use the individual as the unit of interest from the longitudinal point of view. Note, however, that over the duration of the longitudinal survey, the individual may not be available at certain points of time, which may

introduce missing observation. If this happens, the estimating equations should be constructed by taking this missing nature of the data into account. But, in the present report we develop the estimating equations under the assumptions that the longitudinal data is complete. Thus, if the ith individual is assigned design weights w_{is^*} say for his/her inclusion in the sample s^*, then the individual stays in the sample over the whole duration of the longitudinal survey. We now concentrate back to the construction of the estimating equations for β in this case when individuals are units of interest.

In fact the estimating equations in this case follow from the estimating equations (5.6) and (5.7) by considering ϕ as a redundant parameter, and substituting h by i and $H = \sum_{h=1}^{H} n_h$ for $n_h = 1$ by I. This means that the estimating equations developed in the last section reduces to the required estimating equations for this case when household is treated as the individual, and the number of individuals is denoted by I instead of H.

Similar to (5.3), we now write the estimating equations for β as

$$S^*(\beta) = \sum_{i=1}^{I} u_i^*(\beta) = 0, \tag{5.9}$$

and develop a suitable estimator of $S^*(\beta)$ as

$$\hat{S}^*(\beta) = \sum_{i \in s^*} w_{is^*} u_i^*(\beta), \tag{5.10}$$

where w_{is^*} is the survey weight for the ith individual, which may depend on the sample s^*. These equations (5.9) and (5.10) are the same as (4.1) and (4.2), respectively.

Following (5.7), the estimating equations (5.9) may be re-written as

$$S^*(\beta) = \sum_{i=1}^{I} W_i^{*T} \Sigma_{w_i}^{*-1} (Y_i^* - \mu_i^*(\beta)) = 0, \tag{5.11}$$

where Y_i^* is now given by $Y_i^* = (Y_{i1}^{*T}, \ldots, Y_{it}^{*T}, \ldots, Y_{iT_0}^{*T})^T$, with $Y_{it}^* = [Y_{it1}^{*T}, \ldots, Y_{itr}^{*T}, \ldots, Y_{its}^{*T}]^T$, where Y_{itr}^* is the $(J_r - 1)$-dimensional polytomous response vector for the ith individual, corresponding to the rth $(r = 1, \ldots, s)$ variable.

In (5.11), Σ^*_{wi} is the $d^* \times d^*$ working covariance matrix with $d^* = T_0 \sum_{r=1}^{s}(J_r - 1)$. This matrix may be expressed as

$$\Sigma^*_{wi} = A_i^{\frac{1}{2}} C_{wi} A_i^{\frac{1}{2}} \tag{5.12}$$

where $C_{wi} = \oplus D_{it} + \alpha 1_{d^*} 1_{d^*}^T$ is obtained from (5.5) by putting $n_{it} = 1$, that is, $d = d^*$. Here

$$D_{it} = \rho_{it} - \alpha 1_{d_1} 1_{d_1}^T, \tag{5.13}$$

is the $d_1 \times d_1$ matrix with $d_1 = \sum_{r=1}^{s}(J_r - 1)$, as ϕ in (5.6) is now a redundant parameter and d_2 reduces to d_1 as $n_{it} = 1$. Also in (5.12),

$$A_i = \begin{bmatrix} V_{i1}^* & 0 & \cdots & 0 \\ 0 & V_{i2}^* & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & V_{iT_0}^* \end{bmatrix}, \tag{5.14}$$

where V_{it}^* is assumed to be the $d_1 \times d_1$ covariance matrix for the ith individual at time t. The construction of the V_{it}^* matrices for $t = 1, \ldots, T_0$ is shown in the next section.

Now by similar arguments as in the last section, one can use the estimating equation

$$\begin{aligned} S^*(\beta) &= \sum_{i=1}^{I} W_i^{*T} \Sigma_{wi}^{-1} s_i^* \\ &= \sum_{i=1}^{I} \sum_{t=1}^{T_0} W_{it}^{*T} \tilde{s}_{it}^* \\ &= \sum_{i=1}^{I} u_i^*(\beta), \end{aligned} \tag{5.15}$$

yielding

$$u_i^*(\beta) = \sum_{t=1}^{T_0} W_{it}^{*T} \tilde{s^*}_{it},$$

where

$$\tilde{s^*}_{it} = B_{it}^{-1} \left(s_{it}^* - \frac{\alpha}{\eta^*(\alpha)} \sum_{v=1}^{T_0} B_{iv}^{-1} s_{iv}^* \right)$$

with $B_{it} = V_{it}^{*-\frac{1}{2}} D_{it}^{-1}$, $s_{it}^* = y_{it}^* - \mu_{it}^*(\beta)$, and $s_i^* = [s_{i1}^{*T}, \ldots, s_{it}^{*T}, \ldots, s_{iT_0}^{*T}]^T$. In (5.15), W_{it}^{*T} is the $p \times \sum_{r=1}^{s}(J_r - 1)$ matrix obtained from

$$W_i^{*T} = \left(\frac{\partial s_i^{*T}}{\partial \beta}\right)$$

$$= (W_{i1}^{*T}, \ldots, W_{it}^{*T}, \ldots, W_{iT_0}^{*T}).$$

Hence, in order to use the sampling design based estimating equations approach, it is essential to assume that for given β, there exists p-components of a vector $u_i^*(\beta)$ at the finite population level, where the kth component is defined by (5.15). One then includes the $u_i^*(\beta)$ vector in the sample s^* with weight w_{is^*}, for the ith individual, as in (5.10).

5.2 Construction of Marginal Expectation and Covariance Matrix

As mentioned earlier, in this section we concentrate only to the case where individuals are units of interest from the longitudinal point of view. Thus, we discuss the construction of the marginal expectation, $\mu_i^*(\beta)$, and the covariance matrix $\Sigma_{w_i}^*$. Note that the construction of the working covariance matrix $\Sigma_{w_i}^*$ requires only the construction of ρ_{it} matrix in (5.13) and consequently $V_{it}^* = A_{it}$ matrix in (5.14).

5.2.1 Construction of the expectation

Recall from (5.11) that y_i is a $T_0 \sum_{r=1}^{s}(J_r - 1)$-dimensional response vector for the ith individual. That is, $y_i^* = (y_{i1}^{*T}, \ldots, y_{it}^{*T}, \ldots, y_{iT_0}^{*T})$, with $y_{it}^* = [y_{it1}^{*T}, \ldots, y_{itq}^{*T}, \ldots, y_{itr}^{*T}, \ldots, y_{its}^{*T}]^T$, y_{itq}^* being the $(J_q - 1)$-dimensional polytomous response vector corresponding to the qth $(q = 1, \ldots, s)$ variable. The expectation of Y_i^* is denoted by $\mu_i^*(\beta)$ in (5.11), which may be expressed as

$$\mu_i^*(\beta) = (\mu_{i1}^{*T}, \ldots, \mu_{iT_0}^{*T})^T, \tag{5.16}$$

where, for $t = 1, \ldots, T_0$,

$$\mu_{it}^* = (\mu_{it1}^{*T}, \ldots, \mu_{itq}^{*T}, \ldots, \mu_{its}^{*T})^T,$$

with

$$\mu^{*}_{itq} = (\mu_{itr1}, \ldots, \mu_{itqj_q}, \ldots, \mu_{itq(J_q-1)})^T.$$

In the present set-up, Y_{itqj_q} is an indicator variable such that $Y_{itqj_q} = 1$, if the qth ($q = 1, \ldots, s$) response of the ith ($i = 1, \ldots, I$) individual at tth ($t = 1, \ldots, T_0$) time falls into the j_qth ($j_q = 1, \ldots, J_q - 1$) category and zero otherwise.

Note here that although we observe the response y_{itqj_q}, this response is, however, made based on the outcome of an ordinal categorical variable which we will denote by z_{itq}. More specially, for $j_q = 1, \ldots, J_q - 1$,

$$\Pr(y_{itqj_q} = 1) = \Pr(z_{itq} = j_q), \tag{5.17}$$

which, for $j_q = 2, \ldots, J_q$, will be obtained as

$$\Pr(z_{itq} = j_q) = \Pr(z_{itq} \leq j_q) - \Pr(z_{itq} \leq j_q - 1), \tag{5.18}$$

where 'Pr' stands for the probability. The probability given by (5.17) will be denoted by μ_{itqj_q}, and the cumulative probability, $\Pr(z_{itq} \leq j_q)$, will be denoted by $P_{itq}(\ldots, j_q, \ldots)$ which in the present case will be obtained as

$$\begin{aligned} P_{itq}(\ldots, j_q, \ldots) &= \Pr(z_{it1} \leq J_1, \ldots, z_{it(q-1)} \leq J_{q-1}, \\ &\quad z_{itq} \leq j_q, z_{it(q+1)} \leq J_{q+1}, \ldots, z_{its} \leq J_s). \end{aligned} \tag{5.19}$$

For convenience, we denote this cumulative probability in (5.19) by $P_{itq}(j_q)$.

Further note that in order to construct the estimating equations in the multivariate set-up, we will require the correlation matrix for these s variables, which in fact is computed based on all possible collapsed bivariate frequency tables. For the purpose, let us denote the joint cumulative probability for the two variables, say q and r, ($q \neq r$) by $P_{it(qr)}(j_q, j_r)$. That is

$$\begin{aligned} P_{it(qr)}(j_q, j_r) &= \Pr(Z_{it1} \leq J_1, \ldots, Z_{it(q-1)} \leq J_{q-1}, Z_{itq} \leq j_q, \\ &\quad Z_{it(q+1)} \leq J_{q+1}, \ldots, Z_{it(r-1)} \leq J_{r-1}, Z_{itr} \leq j_r, \end{aligned}$$

$$Z_{it(r+1)} \leq J_{r+1}, \ldots, Z_{its} \leq J_s). \tag{5.20}$$

We now concentrate back to the modelling of the marginal probability μ_{itqj_q}. Since the logistic regression is most frequently employed to model the relationship between a binary outcome variable and a set of covariates [cf. Pregibon (1980), Prentice (1976)], after some modifications, we may also use it for modelling the polytomous response variable. Similar to Williamson et al (1995), we consider two types of covariates, say marginal and association covariate. Let X_{itq} denote a p-dimensional possibly time dependent marginal covariate corresponding to the tth ($t = 1, \ldots, T_0$) time and qth ($q = 1, \ldots, 1$) variable for the ith ($i = 1, \ldots, I$) individual or subject. The analogous association covariate will be denoted by $X_{it(a)}$ (say). It is assumed here that the association covariates are not dependent on the variables, rather they depend on the individual and the time when the responses are collected.

We first model the cumulative probability $P_{itq}(j_q)$ for the qth ($q = 1, \ldots, s$) variable by using the polytomous logistic regression

$$\begin{aligned} P_{itq}(j_q) &= \Pr(Z_{itq} \leq j_q) \\ &= \sum_{j'_q=1}^{j_q} \exp\{X_{itq}^T \beta_{qj'_q}\} / \sum_{j_q=1}^{J_q} \exp\{X_{itq}^T \beta_{qj_q}\}, \end{aligned} \tag{5.21}$$

for $j_q = 1, \ldots, J_q$, where β_{qj_q} ($j_q = 1, \ldots, J_q$) is a p-dimensional regression parameter vector. Note that without any loss of generality, we may assume that $\beta_{qJ_q} = 0$, for all $q = 1, \ldots, s$. Now the marginal probabilities or expectations for the qth variable may be written as

$$\begin{aligned} \mu_{itq1} &= P_{itq}(1); \\ \mu_{itqj_q} &= P_{itq}(j_q) - P_{itq}(j_q - 1), \end{aligned}$$

for $j_q = 2, \ldots, J_q - 1$, and

$$\mu_{itqJ_q} = 1 - P_{itq}(J_q - 1), \tag{5.22}$$

where the cumulative probabilities are as in (5.21).

Next let

$$\beta_q = [\beta_{q1}^T, \ldots, \beta_{qj_q}^T, \ldots, \beta_{q(J_q-1)}^T]^T,$$

and

$$\beta = [\beta_1^T, \ldots, \beta_q^T, \ldots, \beta_s^T]^T,$$

where β_q and β are the $p(J_q-1)$ and $p\sum_{q=1}^{s}(J_q-1)$ dimensional parameter vectors, respectively. The purpose of the report is to obtain the estimates for β, say $\hat{\beta}$ and the estimate of the variance of $\hat{\beta}$, say $v(\hat{\beta})$, which we have already given in Sections 4.2 and 4.3.

5.3 Structural covariance (at a given time)

In this section, we construct the covariance matrix of $Y_{it}^* = [Y_{it1}^{*T}, \ldots, Y_{itq}^{*T}, \ldots, Y_{itr}^{*T}, \ldots, Y_{it1}^{*T}]^T$, where Y_{itq}^* $(q = 1, \ldots, s)$ is the (J_q-1)-dimensional random vector corresponding to the qth variable, at time t. Let V_{it}^*, as in (5.14), denote this covariance matrix, i.e.,

$$V_{it}^* = \text{cov}(Y_{it}^*) \tag{5.23}$$

which is a $\sum_{q=1}^{s}(J_q-1) \times \sum_{q=1}^{s}(J_q-1)$ positive definite matrix. For convenience, we first construct the covariance matrix between any two variables, say q and r. Denote this covariance matrix by

$$V_{it(qr)}^* = \text{cov}\begin{pmatrix} Y_{itq}^* \\ Y_{itr}^* \end{pmatrix}$$

$$= \begin{bmatrix} V_{it(qq)} & V_{it(qr)} \\ V_{it(qr)}^T & V_{it(rr)} \end{bmatrix}. \tag{5.24}$$

Here $V_{it(qq)}$ is the $(J_q-1) \times (J_q-1)$ covariance matrix for the qth variable, which is given by

$$V_{it(qq)} = \text{diag}[\mu_{itq1}, \ldots, \mu_{itqj_q}, \ldots, \mu_{itq(J_q-1)}]$$

$$-\mu_{itq}^* \mu_{itq}^{*T}, \tag{5.25}$$

where $\mu^*_{itq} = [\mu_{itq1}, \ldots, \mu_{itq(J_q-1)}]^T$ is the $(J_q - 1)$-dimensional marginal expectations vector with its components as in (5.22).

Note that as opposed to $V_{it(qq)}$, the construction of the $V_{it(qr)}$ matrix, for $q \neq r$, is not easy. This is because, these covariance computations require an extra modelling for the association between two dichotomized variables. More specifically, we write the (j_q, j_r)th element of the covariance matrix $V_{it(qr)}$ as

$$v_{it(qr)} = \xi_{itj_qj_r} - \mu_{itqj_q}\mu_{itrj_r}, \tag{5.26}$$

where

$$\begin{aligned} \xi_{itj_qj_r} &= \Pr(Z_{itq} = j_q, Z_{itr} = j_r) \\ &= P_{it(qr)}(j_q, j_r) + P_{it(qr)}(j_q - 1, j_r - 1) \\ &\quad - P_{it(qr)}(j_q, j_r - 1) - P_{it(qr)}(j_q - 1, j_r), \end{aligned} \tag{5.27}$$

by (5.20). Now the modelling of the cumulative probability $P_{it(qr)}(j_q, j_r)$ requires the correlation structure between a pair of correlated binary variables, say T_{itqj_q} and T_{itrj_r}, be known. These correlated binary variables are defined such that

$$\begin{aligned} &\Pr(T_{itqj_q} = 1, T_{itrj_r} - 1) \\ &= \Pr(Z_{itq} \leq j_q, Z_{itr} \leq j_r) \\ &= P_{it(qr)}(j_q, j_r), \end{aligned}$$

which are easy to interpret based on the 2×2 contingency table [cf. Molenberghs and Lasaffre (1994)]

		T_itrj_r	
		1	0
T_{itqj_q}	1	$Z_{itq} \leq j_q, Z_{itr} \leq j_r$	$Z_{itq} \leq j_q, Z_{itr} > j_r$
	0	$Z_{itq} > j, Z_{itr} \leq j_r$	$Z_{itq} > j_q, Z_{itr} > j_r$

(5.28)

obtained by dichotomizing the $J_q \times J_r$ contingency table at (j_q, j_r), $j_q = 1, \ldots J_q - 1$, $j_r = 1, \ldots, J_r - 1$. For this $q \neq r$ case, cumulative probabilities corresponding to the cells in (5.28) are

$P_{it(qr)}(j_q, j_r)$	$P_{itq}(j_q) - P_{it(qr)}(j_q, j_r)$
$P_{itr}(j_r) - P_{it(qr)}(j_q, j_r)$	$1 - P_{itq}(j_q) - P_{itr}(j_r) + P_{it(qr)}(j_q, j_r)$

(5.29)

where $P_{itq}(j_q) = \sum_{j_r=1}^{J_r} P_{it(qr)}(j_q, j_r)$, and $P_{itr}(j_r) = \sum_{j_q=1}^{J_q} P_{it(qr)}(j_q, j_r)$. Many authors [cf. Dale (1986), Molenberghs and Lesaffre (1994), Williamson et al (1995)] have modelled the association between T_{itqj_q} and T_{itrj_r} by using the global odds ratios and then computed the joint cumulative probabilities $P_{it(qr)}(j_q, j_r)$ based on known global odds ratios. When global odds ratios are unknown, which is usually the case, they are estimated through an additional suitable model. In the present approach, unlike these authors, we model the association between two variables by using the well-known Pearsonian type correlations. We consider two cases. First, under the assumption that the correlation structure for the appropriate dichotomized variables remain the same for all individuals $i = 1, \ldots, I$. We also consider the case when the correlations structures may be different. In the latter case, we model the correlations by using two approaches as discussed in Section 5.3.2.

In general, it follows from (5.28) and (5.29) that the bivariate cumulative probability $P_{it(qr)}(j_q, j_r)$ in (5.27) may be computed by using its relationship with the Pearsonian correlation, $\rho^*_{itj_qj_r}$, between the dichotomized variables T_{itqj_q} and T_{itrj_r}. More specifically, $P_{it(qr)}(j_q, j_r)$ may be obtained from

$$\rho^*_{itj_qj_r} = \frac{P_{it(qr)}(j_q, j_r) - P_{itq}(j_q)P_{itr}(j_r)}{\left[\{P_{itq}(j_q)(1 - P_{itq}(j_q)\}^{\frac{1}{2}}\{P_{itr}(j_r)(1 - P_{itr}(j_r)\}^{\frac{1}{2}}\right]}, \tag{5.30}$$

where, for $j_q = 1, \ldots, J_q - 1$; $j_r = 1, \ldots, J_r - 1$, $P_{itq}(j_q)$ and $P_{itr}(j_r)$ are the marginal cumulative probabilities given by (5.21). For known β, these marginal cumulative probabilities are known. Therefore, to know the bivariate cumulative probability, one needs to know the correlation structure through $\rho^*_{itj_qj_r}$. We consider the following two cases.

5.3.1 Identical correlation structure for individuals

Note that the correlations in (5.30) between the two correlated binary variables appear to vary among individuals ($i = 1, \ldots, I$). It is, however, common in practice to assume that these correlations

remain the same for all independent individuals. For example, we refer to the cluster regression analysis for the repeated discrete or continuous data by Liang and Zeger (1986), and a more recent study by Lipsitz and Fitzmaurice (1996). In these studies, the association between the repeated binary responses, for example, are considered to be the same for all individuals, although their means and variances are generally different for individuals. In this section, in the spirit of Liang and Zeger, Lipsitz and Fitzmaurice, we assume that the correlations of the dichotomized variables remain the same for all I individuals. Consequently, we may pool the information from all I individuals and estimate the common correlations by using the formula

$$\hat{\rho}^{*}_{tj_qj_r} = \frac{1}{I}\sum_{i=1}^{I} \frac{I(T_{itqj_q}=1, T_{itrj_r}=1) - \hat{P}_{itq}(j_q)\hat{P}_{itr}(j_r)}{\left[\{\hat{P}_{itq}(j_q)(1-\hat{P}_{itq}(j_q)\}^{\frac{1}{2}}\{P_{itr}(j_r)(1-P_{itr}(j_r)\}^{\frac{1}{2}}\right]} \tag{5.31}$$

where $\hat{P}_{itq}(j_q)$ and $\hat{P}_{itr}(j_r)$ are the estimates of their respective cumulative probabilities, which are computed by using $\hat{\beta}$ for β in $P_{itq}(j_q)$ and $P_{itr}(j_r)$, $\hat{\beta}$ being a suitable estimate of β.

Further note that since $E[I(T_{itqj_q}=1, T_{itrj_r}=1)] = P_{it(qr)}(j_u, j_r)$, it follows from (5.31) that for large I, $\hat{\rho}_{tj_qj_u}$ are consistent estimate of $\rho_{tj_qj_r}$, provided $\hat{\beta}$ is a consistent estimate for β. The consistency of $\hat{\beta}$ for β was discussed in Section 4.

We also note that for $j_q = 1, \ldots, J_q - 1$, $j_r = 1, \ldots, J_r - 1$, $\hat{\rho}^{*}_{tj_qj_r}$ in (5.31) should satisfy the restriction

$$\hat{L}^{0}_{tj_qj_r} < \hat{\rho}^{*}_{tj_qj_r} < \hat{U}^{0}_{tj_qj_r} \tag{5.32}$$

where

$$\begin{aligned}\hat{L}^{0}_{tj_qj_r} &= \max\Big[-\{(\hat{P}_{itq}(j_q)\hat{P}_{itr}(j_r))/(\hat{Q}_{itq}(j_q)\hat{Q}_{itr}(j_r))\}^{\frac{1}{2}}, \\ &\quad -\{(\hat{Q}_{itq}(j_q)\hat{Q}_{itr}(j_r))/(\hat{P}_{itq}(j_q)\hat{P}_{itr}(j_r))\}^{\frac{1}{2}}\Big],\end{aligned}$$

and

$$\hat{U}^{0}_{tj_qj_r} = \min\Big[\{\hat{P}_{itr}(j_r)\hat{Q}_{itq}(j_q)/\hat{P}_{itq}(j_q)\hat{Q}_{itr}(j_r)\}^{\frac{1}{2}},$$

$$\{\hat{P}_{itq}(j_q)/\hat{P}_{itr}(j_r)\hat{Q}_{itq}(j_q)\}^{\frac{1}{2}}\Big]$$

with $\hat{Q}_{itq}(j_q) = 1 - \hat{P}_{itq}(j_q)$, and $\hat{Q}_{itr}(j_r) = 1 - \hat{P}_{itr}(j_r)$. Note that these restrictions are necessary to have the covariance matrix under construction as a positive definite matrix.

Now by using $\hat{\rho}^*_{(j_q j_r)}$ from (5.31) for $\rho^*_{it j_q j_r}$ in (5.30), we compute the estimate of $P_{itqr}(j_q, j_r)$ as

$$\begin{aligned}\hat{P}_{itqr}(j_q, j_r) &= \hat{P}_{itq}(j_q)\hat{P}_{itr}(j_r) + \hat{\rho}^*_{(j_q j_r)}\Big[\{P_{itq}(j_q)(1 - P_{itq}(j_q))\}^{\frac{1}{2}} \\ &\quad \times \{\hat{P}_{itr}(j_r)(1 - \hat{P}_{itr}(j_r))\}^{\frac{1}{2}}\Big], \end{aligned} \tag{5.33}$$

yielding $\hat{\xi}_{it j_q j_r}$ by (5.27) and the estimate of covariance in (5.26) as

$$\hat{v}_{it(qr)}(j_q, j_r) = \hat{\xi}_{it j_q j_r} - \hat{\mu}_{itq j_q}\hat{\mu}_{itr j_r}. \tag{5.34}$$

We have thus constructed a suitable estimate for the covariance matrix $V_{it(qr)}$ in (5.24). By using this $V_{it(qr)}$ and $V_{it(qq)}$ from (5.25), one may then compute $V^*_{it(qr)}$ by (5.24), which is the covariance matrix of $[Y^{*T}_{itq}, Y^{*T}_{itr}]^T$. By (5.25) and (5.34), one may however directly write the covariance matrix $V^*_{it} = \text{cov}(Y^*_{it})$ (5.23) as

$$V^*_{it} = \begin{bmatrix} V_{it(11)} & V_{it(12)} & \cdots & V_{it(1s)} \\ & V_{it(22)} & \cdots & V_{it(2s)} \\ & & \ddots & \vdots \\ & & & V_{it(ss)} \end{bmatrix}. \tag{5.35}$$

5.3.2 Variable correlation structure for individuals

For the cases when the correlation structures vary for the individuals, one may attempt to model the correlation structures in different ways. We discuss two approaches below.

Approach 1. In this approach, first, similar to Williamson et al (1995), we separate the association covariates from the marginal covariates. As mentioned in Section 5.2.1, the association covariates will be denoted by $X_{it(a)}$. That is, $X_{it(a)}$ is the possible time dependent association covariate vector

of dimension p^*, corresponding to the time t for the ith individual. Next, similar to Darlington (1992) , one may model the correlations of the standardized residuals of the two dichotomized variables as

$$\rho^{**}_{it j_q j_r} = \exp\{X^T_{it(a)}\eta\}/[1 + \exp\{X^T_{it(a)}\eta\}], \tag{5.36}$$

which is a simple logistic representation of the correlation with dependence on covariates that vary from individual to individual, and possibly from time to time. This modelling, therefore, does not allow any negative correlations between the residuals of the dichotomized variables T_{itqj_q} and T_{itrj_r}, explained in the previous section. In (5.36), η is a p^*-dimensional parameter vector. Note, however, that the marginal and cumulative probabilities are defined as before in terms of the marginal covariate X_{itq}. For known β, the correlation of the residuals in (5.36) is the same as the correlation of the dichotomized variables in (5.30). Consequently, by using (5.36) in (5.30), the bivariate cumulative probability reduces to

$$\begin{aligned} P^*_{it(qr)}(j_q, j_r|\eta) &= P_{itq}(j_q)P_{itr}(j_r) \\ &\quad + \{\exp(X^T_{it(a)}\eta)\}\{1 + \exp(X^T_{it(a)}\eta)\}^{-1} \\ &\quad \times \{P_{itq}(j_q)(1 - P_{itq}(j_q))P_{itr}(j_r)(1 - P_{itr}(j_r))\}^{\frac{1}{2}}, \end{aligned} \tag{5.37}$$

yielding the joint probability

$$\begin{aligned} \xi^*_{itqr}(j_q, j_r|\eta) &= P^*_{it(qr)}(j_q, j_r|\eta) + P^*_{it(qr)}(j_q - 1, j_r - 1|\eta) \\ &\quad - P^*_{it(qr)}(j_q - 1, j_r|\eta) - P^*_{it(qr)}(j_q, j_r - 1|\eta). \end{aligned} \tag{5.38}$$

At this stage, we are interested to estimate η only. Now for known $\beta = \hat{\beta}$ (say), we develop the estimating equation for η as follows. Let $U_{itqr}(j_q, j_r) = I(y_{itqj_q} = 1, y_{itrj_r} = 1)$ be an indicator variable. It then follows that

$$E\{U_{itqr}(j_q, j_r)\} = \xi^*_{itqr}(j_q, j_r|\eta)$$

and

$$\text{var}\{U_{itqr}(j_q, j_r)\} = \xi^*_{itqr}(j_q, j_r|\eta)\{1 - \xi^*_{itqr}(j_q, j_r)\}. \tag{5.39}$$

Next, define

$$U^*_{it} = [U^{*T}_{it12}, \ldots, U^{*T}_{itqr}, \ldots, U^{*T}_{it(s-1)s}]^T, \tag{5.40}$$

where, for $q \neq r$,

$$U^*_{itqr} = [U_{itqr}(1,1), \ldots, U_{itqr}(j_u, j_r), \ldots, U_{itqr}(J_q - 1, J_r - 1)]^T$$

is a $(J_q - 1)(J_r - 1)$-dimensional vector of indicator variables. Here U^*_{it} is the $\sum_{q \neq r}^{s}(J_q - 1)(J_r - 1) \times 1$ vector of unit vectors. It then follows that

$$E(U^*_{it}) = \xi^*_{it} = [\xi^{*T}_{it12}, \ldots, \xi^{*T}_{itqr}, \ldots, \xi^{*T}_{it(s-1)s}]^T, \tag{5.41}$$

with

$$\xi^*_{itqr} = [\xi^*_{itqr}(1,1|\eta), \ldots, \xi^*_{itqr}(j_q, j_r|\eta), \ldots, \xi^*_{itqr}(J_q - 1, J_r - 1|\eta)].$$

Now, by pretending that the indicator variables are independent, we construct a working covariance matrix of U^*_{it} given by

$$\begin{aligned} R^*_{it} &= \text{cov}(U^*_{it}) \\ &= \text{diag}[\ell_{it11}(1,1|\eta)m_{it11}(1,1|\eta), \ldots, \ell_{itqr}(j_q, j_r|\eta)m_{itqr}(j_q, j_r|\eta), \ldots \\ &\quad \ldots, \ell_{it(s-1)s}(J_{s-1} - 1, J_s - 1|\eta)m_{it(s-1)s}(J_{s-1} - 1, J_s - 1|\eta)], \end{aligned} \tag{5.42}$$

where $\ell_{itqr}(j_q, j_r|\eta) = \xi^*_{itqr}(j_q, j_r|\eta)$, and $m_{itqr}(j_q, j_r|\eta) = 1 - \ell_{itqr}(j_q, j_r|\eta)$.

By combining (5.40), (5.41), and (5.42), we then construct the estimating equations for η given by

$$\sum_{i=1}^{I} C^T_{it} R^{*^{-1}}_{it}(U^*_{it} - \xi^*_{it}) = 0, \tag{5.43}$$

where $C_{it} = \partial\xi^*_{it}/\partial\eta$. The solution of (5.43), denoted by $\hat{\eta}$, may be obtained by the customary Newton-Raphson method. Given the values $\hat{\eta}(u_0)$ at the uth iteration, $\hat{\eta}(u_0+1)$ is obtained as

$$\hat{\eta}(u_0+1) = \hat{\eta}(u_0) + \left[\sum_{i=1}^{I} C_{it}^T R_{it}^{*^{-1}} C_{it}\right]^{-1}_{u_0} \left[\sum_{i=1}^{I} C_{it}^T R_{it}^* (U_{it}^* - \xi_{it}^*)\right]_{u_0}, \tag{5.44}$$

where $[\ ]_{u_0}$ denotes that the expression within the brackets is evaluated at $\hat{\eta}(u_0)$.

Next the estimate $\hat{\eta}$ is used in (5.37) and (5.38) to compute the (j_q, j_r)th element of the covariance matrix $V_{it(qr)}$, as

$$\tilde{v}_{it(qr)}(j_q, j_r) = \xi^*_{itqr}(j_q, j_r|\hat{\eta}) - \hat{\mu}_{itqj_q}\hat{\mu}_{itrj_r}, \tag{5.45}$$

where $\hat{\mu}_{itqj_q}$ and $\hat{\mu}_{itrj_r}$ are as in (5.34). Consequently, by (5.24) we obtain the covariance matrix of $[Y_{itq}^{*^T}, Y_{itr}^{*^T}]^T$ given by

$$\tilde{V}^*_{it(qr)} = \begin{bmatrix} V_{it(qq)} & \tilde{V}_{it)qr)} \\ \tilde{V}^T_{it(qr)} & V_{it(rr)} \end{bmatrix}, \tag{5.46}$$

yielding the covariance matrix of Y^*_{it} as

$$\tilde{V}^*_{it} = \begin{bmatrix} V_{it(11)} & \tilde{V}_{it(12)} & \cdots & \tilde{V}_{it(1s)} \\ & V_{it(22)} & \cdots & \tilde{V}_{it(2s)} \\ & & \ddots & \vdots \\ & & & V_{it(ss)} \end{bmatrix}. \tag{5.47}$$

Approach 2. A Mixed Effects Approach

In the last section we have modelled the correlations between two variables q and r for the ith individual so that they vary from individual to individual through some association covariates. But as it is seen from (5.36) that this modelling does not allow any negative values for the correlations of the dichotomized residuals. As far as the range for correlation is concerned, there is, however, no problem with the construction of the correlation structure in Section 5.3.1, except that it may be a strong assumption to consider identical correlation for all individual in the study. To overcome these two above problems, in this section, we propose a mixed effects approach where the correlations between two ordinal variables will vary from individual to individual and they may be positive or

negative as desired. The specific modelling is discussed below. Note that this approach described here is an alternative approach only to construct a working correlation matrix.

The proposed modelling in fact will be an extension of the modelling for the identical correlations structure discussed in Section 5.3.1. More specifically, we undertake a simple extension so that when the variance component of the random effects is zero, the present model will reduce to the model discussed in approach 1. But, in the positive variance component case, we will adopt an adhoc estimation approach, where the identical correlation obtained in Section 5.3.1 will be adjusted for the positive variance of the random effects which can vary from individual to individual.

Let ϵ_i be a latent random variable such that for given ϵ_i, the marginal cumulative probabilities for the qth variable are given by

$$\begin{aligned} \tilde{P}_{itq}(j_q) &= \Pr(Z_{itq} \leq j_q) \\ &= \epsilon_i \sum_{j'_q=1}^{j_q} \exp\{X_{itq}^T \beta_{qj'_q}\} / \sum_{j_q=1}^{J_q} \exp\{X_{itq}^T \beta_{qj_q}\} \end{aligned} \tag{5.48}$$

for $j_q = 1, \ldots, J_q - 1$, with $P_{itq}(J_q) = 1$. Consequently, the conditional marginal probability for the qth variable may be expressed as

$$\begin{aligned} \tilde{\mu}_{itq1} &= \tilde{P}_{itq}(1) = \epsilon_i P_{itq}(1) \\ \tilde{\mu}_{itqj_q} &= \epsilon_i \{P_{itq}(j_q) - P_{itq}(j_q - 1)\}, \end{aligned}$$

for $j_q = 2, \ldots, J_q - 1$, and

$$\tilde{\mu}_{itqJ_q} = 1 - \epsilon_i P_{itq}(J_q - 1) \tag{5.49}$$

where $P_{itq}(j_q) = \sum_{j'_q=1}^{j_q} \exp\{X_{itq}^T \beta_{qj'_q}\} / \sum_{j_q=1}^{J_q} \exp\{X_{itq}^T \beta_{qj_q}\}$ as in (5.22).

Suppose that $\epsilon_i \sim (1, \sigma_\epsilon^2 b_i(\eta))$, where, for example, $b_i(\eta) = \exp\{X_{it(a)}^T \eta\}[1 + \exp\{X_{it(a)}^T \eta\}]^{-1}$, which is the same as the variable correlation itself introduced in approach 1. Also suppose that ϵ_i's are independent and σ_ϵ^2 is a small unknown non-negative quantity so that $E\epsilon_i^r$ is of $o(\sigma_\epsilon^2)$ for $r \geq 3$.

Next suppose that in the $J_q \times J_r$ contingency table for the qth and rth variables, the joint bivariate cumulative probabilities are simply ϵ_i multiple of the joint cumulative probabilities. This yields the four cell probabilities for the dichotomized variables T_{itqj_q} and T_{itrj_r} defined in (5.28), as

$\epsilon_i P_{it(qr)}(j_q, j_r)$	$\epsilon_i\{P_{itq}(j_q) - P_{it(qr)}(j_q, j_r)\}$
$\epsilon_i\{P_{itq}(j_r) - P_{it(qr)}(j_q, j_r)\}$	$1 - \epsilon_i P_{itq}(j_q) - \epsilon_i P_{itq}(j_r) + \epsilon_i P_{it(qr)}(j_q, j_r)$

(5.50)

The dichotomized probability table in (5.50) is quite similar to that of (5.29). It then follows that given ϵ_i, the correlation between the dichotomized variables T_{itqj_q} and T_{itrj_r} for the ith individual is given by

$$\tilde{\rho}_{itj_qj_r}(\epsilon_i) = \frac{P_{it(qr)}(j_q, j_r) - \epsilon_i P_{itq}(j_q) P_{itr}(j_r)}{[\{P_{itq}(j_q)\{1 - \epsilon_i P_{itq}(j_q)\}P_{itr}(j_r)\{1 - \epsilon_i P_{itq}(j_r)\}]^{\frac{1}{2}}}. \tag{5.51}$$

Note that all the marginal and cumulative probabilities modelled in this section reduce to those in Section 5.3, when the former probabilities are evaluated at $\epsilon_i = E(\epsilon_i) = 1$. This is also true for the correlation defined in (5.51) as it reduces to (5.30) when evaluated at $\epsilon_i = 1$. These two correlations may be referred to as the correlation between dichotomized variables for the ith person under mixed and fixed effects models, respectively.

Let $\tilde{\rho}'_{itj_qj_r} = \partial\tilde{\rho}_{itj_qj_r}/\partial\epsilon_i$, and $\tilde{\rho}''_{itj_qj_r} = \partial^2\rho_{itj_qj_r}/\partial\epsilon_i^2$. Then upto $o(\sigma_\epsilon^2)$, the correlation in (5.51) may be expressed as

$$\tilde{\rho}_{itj_qj_r} = \rho^*_{itj_qj_r} + \{\sigma_\epsilon^2 b_i(\eta)/2\}[\tilde{\rho}''_{itj_qj_r}]_1, \tag{5.52}$$

where $[\,]_1$ denotes that the expression within the bracket is evaluated at $\epsilon_i = 1$.

Suppose we now assume that the variable correlations among the individuals occur only through the variance of the random effects ϵ_i. It is then reasonable to assume that $\rho^*_{itj_qj_r}$ remains the same for all individuals $i = 1, \ldots, I$. Consequently, we write

$$\tilde{\rho}_{itj_qj_r}(\psi) = \rho^*_{tj_qj_r} + \{\sigma_\epsilon^2 b_i(\eta)/2\}[\tilde{\rho}''_{tj_qj_r}]_1, \tag{5.53}$$

where $\psi = (\eta^T, \sigma_\epsilon^2)^T$. Next for known β, ψ may be estimated in the manner similar to that for η in (5.43). For this, we first compute the unconditional joint cumulative probability by

$$P_{it(qr)}(j_q, j_r|\psi) \quad = \quad P_{itq}(j_q) P_{itr}(j_r)$$

$$+\tilde{\rho}_{it_{j_q j_r}}(\psi)[P_{itq}(j_q)\{1-P_{itq}(j_q)\}P_{itr}(j_r)\{1-P_{itr}(j_r)\}]^{\frac{1}{2}}, \qquad (5.54)$$

which yields

$$\begin{aligned}\xi_{itqr}(j_q, j_r|\psi) &= P_{it(qr)}(j_q, j_r|\psi) + P_{it(qr)}(j_q-1, j_r-1|\psi) \\ &\quad - P_{it(qr)}(j_q-1, j_r) - P_{it(qr)}(j_q, j_r-1|\psi). \qquad (5.55)\end{aligned}$$

Consequently, by using $\xi_{itqr}(j_q, j_r|\psi)$ in an estimating equation similar to (5.43) we can obtain the estimate of ψ, say $\hat{\psi}$, which provides the covariance matrix of $[Y_{itq}^T, Y_{itr}^T]^T$, say $\tilde{\tilde{V}}_{it}$. The computation of $\tilde{\tilde{V}}_{it}$ is quite similar to that of $\tilde{V}_{it}^*$ in (5.47). This $\tilde{\tilde{V}}_{it}$ matrix, will be used to update the estimate of β as in the next section.

5.4 Use of the Mean Vector and the Covariance Matrix in Estimating β

Recall from Section 5.2.1 that

$$E(Y_i^*) = \mu_i^*(\beta), \qquad (5.56)$$

as in (5.16), where $Y_i^* = (Y_{i1}^{*T}, \ldots, Y_{it}^{*T}, \ldots, Y_{iT_0}^{*T})^T$. Here Y_{it}^* is the $\sum_{q=1}^{s}(J_q-1)$-dimensional response vector collected at time t, for all s ordinal variables. Now, the covariance matrix of Y_{it}^* has been computed under three different situations. In Section 5.3.1, the covariance matrix of Y_{it} has been computed as V_{it}^* by (5.35), under the identical correlation structure for all I individuals. Note that this correlation structure is actually meant for the appropriate dichotomized variables involved in computing V_{it}^*. In Section 5.3.2, $\text{cov}(Y_{it}^*) = \tilde{V}_{it}^*$, is computed based on a variable correlation structure for the dichotomized variables involved and modelling the correlations through a fixed effects model. Finally, mixed effects approach has been used and covariance of Y_{it}^* has been computed similar to $\tilde{V}_{it}^*$. This new covariance has been denoted by $\tilde{\tilde{V}}_{it}$.

We can now compute the correlation matrix ρ_{it} based on V_{it}^*, $\tilde{V}_{it}^*$ or $\tilde{\tilde{V}}_{it}$. This ρ_{it} is a $\sum_{q=1}^{s}(J_q - 1) \times \sum_{q=1}^{s}(J_q-1)$ symmetric matrix. By using this matrix in (5.13), we compute the D_{it} matrix as a

function of α, the longitudinal correlation, and then compute the $\Sigma^*_{u_i}$ by (5.12). Finally, the sample estimator, $\hat{\beta}$, is obtained by solving $\hat{S}^*(\hat{\beta}) = 0$, as in Section 4, where

$$\begin{aligned} \hat{S}^*(\beta) &= \sum_{i \epsilon s^*} w_{is^*} u_i^* \beta \\ &= \sum_{i \epsilon s^*} w_{is^*} W_i^{*T} \Sigma_{u_i}^{*-1} (y_i^* - \mu_i^*(\beta)) \end{aligned} \tag{5.57}$$

as in (4.4). The asymptotic properties of the $\hat{\beta}$ estimator is given by (4.7) in Section 4.3.

When α is unknown, we follow Quenouille (1958) and obtain a pooled estimate of the longitudinal correlation as

$$\hat{\alpha} = \frac{\sum_{i \epsilon s^*} \sum_{r=1}^{s} \sum_{t \neq t'}^{T_0} w_{is^*} r_{itr}^* r_{it'r}^*}{\sum_{i \epsilon s^*} \sum_{r=1}^{s} \sum_{t=1}^{T_0} w_{is^*} r_{itr}^{*2}} \tag{5.58}$$

where $r_{itr}^* = (y_{itr}^* - \mu_{itr}^*)/\{var(y_{itr}^*)\}^{\frac{1}{2}}$, and $var(y_{itr}^*)$ is the variance of Y_{itr}^* obtained from the V_{it}^* matrix given in (5.23). When the complex survey design is known, the estimate of this longitudinal correlation parameter may be computed based on the specific nature of the design (such as stratified cluster sampling).

6 Remarks and Further Investigation

Analyzing longitudinal survey data is an extremely important topic. One of the main objectives of such longitudinal analysis is to determine the changes that take place in households and families or individuals over time. Unfortunately, it is, however, not easy to analyze this types of data. Some of the issues that complicate the methodological development in this area are : (1) the responses of an individual collected over a period of time are correlated; (2) multi-dimensional responses make this issue more complicated as then there will be a structure correlation matrix for the variables concerned; (3) missing longitudinal responses; (4) variable household composition over the period of the survey; (5) the use of complex survey along with the expectation of cross-sectional estimation.

In this report, we have concentrated to the longtudinal estimation as opposed to the cross-sectional estimation. We have developed a multivariate regression approach which takes the structural as well as longitudinal correlations into account. It has been assumed that there is no missing information over the period of time.

The methodology developed in this report is an important step toward the analysis of longitudinal survey data, such as SLID data collected by Statistics Canada. In the next report, we will analyse this data set using the present methodology. The methodology will further be modified to incorporate the longitudinal missing responses. An attempt will also be made to deal with variable household composition, which will be useful for the cross-sectional estimation at any given year.

Acknowledgement

The author thanks Dr. A. C. Singh for some useful suggestions leading to the improvement of the report.

References

Binder, D. A. (1983). On the variance of asymptotically normal estimators from complex surveys, *Int. Statist. Rev.* 51, 279-292.

Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses, *Biometrics* 42, 909-917.

Darlington, G. A. (1992). Analysis of binary longitudinal data, *In Proceedings of Statistics Canada Symposium*; Design and Analysis of Longitudinal Surveys, 133-139.

Fuller, W. A. (1975). Regression analysis for sample surveys, *Sankhya* C37, 117-132.

Hapuarachci, K. P. (1996). Analysis of longitudinal data, *Statistics Canada Report.*

Latouche, M., and Michaud, S. (1995). Cross-section weighting of a longitudinal surveys and its impact on analysis, Statistics Canada Report.

Lavallee, P. , and Hunter, L. (1992). Weighting for the survey of labour and income dynamics, *In Proceedings of Statistics Canada Symposium*: Design and Analysis of Longitudinal Surveys, 65-75.

Lawless, J. F. (1997). Analysis of event history data. *Statistics Canada Report.*

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

Lipsitz, S. R., and Fitzmaurice, G. M. (1996). Estimating equations for measures of association between repeated binary responses, *Biometrics* 52, 903-912.

Molenberghs, G., and Lesaffre, E. (1994). Marginal modelling of correlated ordinal dtaa using a multivariate Plackett distribution, *J. Amer. Statist. Assoc.* 89, 633-644.

Pregibon, D. (1980). Goodness-of-link tests for generalized linear models, *Applied Statistics* 29, 15-24.

Prentice, R. L. (1976). A generalization of the probit and logit methods for dose response curves, *Biometrics* 32, 761-768.

Quenouille, M. H. (1958). The comparison of correlations in time series, *Journal of Royal Statistical Society.* B, 20, 158-168.

Rao, J. N. K., Scott, A. J., and Skinner, C. J. (1997). Quasi-score tests with survey data, *preprint.*

Tambay, J.-L., and Catlin, G. (1995). Sample design of the national population health survey, *Health Reports, Statistics Canada, Cat. No.* 82-003.

Williamson, J. M., Kim, K., and Lipsitz, S. R. (1995). Analyzing bivariate ordinal data using a global odds ratios, *J. Amer. Statist. Assoc.* 90, 1432-1437.