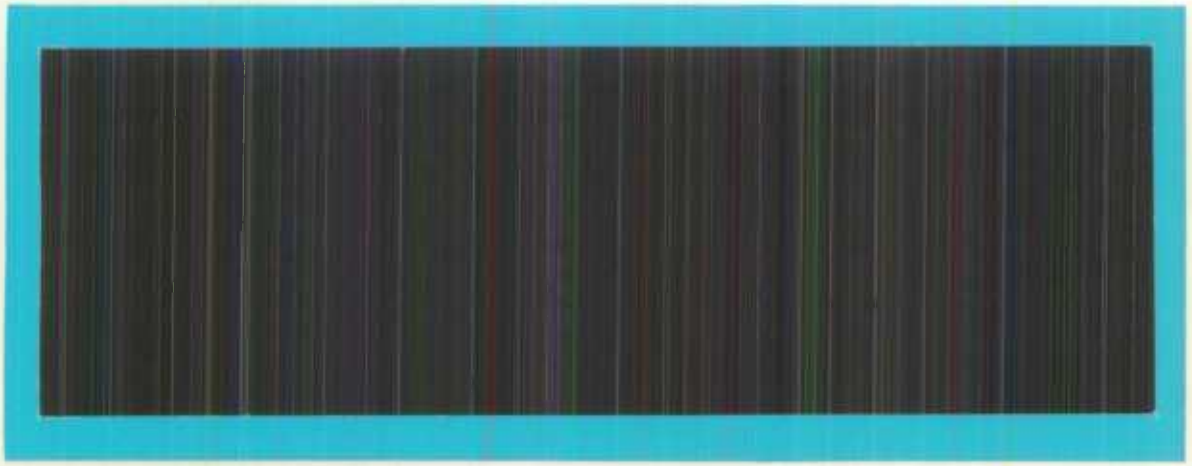
Statistics Canada Statistique Canada



Methodology Branch

Social Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes sociales

Canada

WORKING PAPER

METHODOLOGY BRANCH

REGRESSION ANALYSIS FOR ORDINAL LONGITUDINAL SURVEY DATA WITH OUTCOME SUBJECT TO NONRESPONSE: PART II

SSMD - 99-006 E

Brajendra C. Sutradhar

Memorial University of Newfoundland

A Report Prepared for Statistics Canada

Revised: March 9, 1999

Abstract

One of the main objectives of longitudinal analysis is to determine the changes that take place in households and families or individuals over time. One is interested to examine the effects of the associated covariates on one or more suitable response variables, after taking the longitudinal correlations into account. In this paper, a multivariate approach was developed to analyse ordinal polytomous longitudinal survey data, such as SLID data collected by Statistics Canada. This methodology has been applied to the SLID data to examine the effects of covariates such as age sex, marital status, race, education level and province of residence on the "jobless spell" response variable. The longitudinal correlations between responses for 1993 and 1994 were modelled following Lipsitz and Fitzmaurice (1996) and estimated following Sutradhar and Das (1999).

Résumé

L'un des principaux objectifs de l'analyse longitudinale est de déterminer les changements survenus avec le temps dans les ménages et familles ou chez les individus. Une telle analyse examine les effets des covariables associées à une ou plusieurs variables de réponse pertinentes, après avoir pris en compte les corrélations longitudinales. Dans le présent document, une approche multivariable a été 'elaborée pour analyser les données polytomiques ordinales d'enquêtes longitudinales, par exemple les données de l'EDTR recueillies par Statistique Canada. Cette méthodologie a été appliquée aux données de l'EDTR afin d'examiner les effets de covariables comme l'ge, le sexe, la situation de famille, l'origine ethnique, la scolarité et la province de résidence sur la variable de réponse sans emploi . Les corrélations longitudinales entre les réponses de 1993 et 1994 ont été modélisées sur le travail de Lipsitz et Fitzmaurice (1996) et estimées conformément à celui de Sutradhar et Das (1999).

CONTENTS

1 A Brief Review of Part I

Analysing longitudinal survey data is an important topic. One of the main objectives of such longitudinal analysis is to determine changes that take place in households and families or individuals over time. More precisely, one may like to describe the marginal expectation of the outcome variable as a function of the covariates while accounting for the cross-sectional as well as longitudinal correlations. To meet this objective, I developed a multivariate regression approach in the first part of the study, which takes the structural as well as longitudinal correlations into account. This theoretical development was motivated by the needs for methodologies to analyse longitudinal survey data, such as SLID data, collected by Statistics Canada. As the responses under longitudinal study can be categorized into more than two ordinal groups, in part I, each of the response variables in the multi-dimensional set-up was assumed to follow a suitable multinomial distribution with suitable cumulative logistic probabilities reflecting the ordinal nature of the responses. One of the main objectives of the present report is to apply the theoretical findings of part I to a Statistics Canada longitudinal data set, namely SLID. The results of the application are provided in detail in Section 3.

Note, however, that the regression methodologies in part I were developed under the assumption that the longitudinal survey data is complete, that is, there is no missing data. Consequently, in Section 3 of the current report, we have applied the methodologies to those individuals only who provided the complete information (16890 in this case) over the duration of the survey (first two waves). But, there are possibilities where some of the individuals may be missing from the survey occassionally or forever, or the individual does not provide complete information. This problem of missing responses was not dealt in the first report, which we discuss now in detail in Section 4.

Finally, we conclude the report in Section 5 by making some comments on other related issues those are thought to be useful to address in modelling and analysing any longitudinal survey data similar to that analyzed in this report. These issues, for example, are: robustness of the longitudinal correlations structure, robustness of the regression estimators under model misspecification, goodness

of fit of the model in general.

2 SLID Data: An Exploratory Analysis

The survey of Labour and Income Dynamics (SLID) 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 overlapping 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. The second panel was introduced in 1997. Every year, information is collected on the panel members' labour market activity and income during the preceding year. Thus, for the first panel, information for 1993, 1994 and 1995 were collected in full in January of 1994, 1995 and 1996 respectively. In this problem, it may be of interest to determine the causes of movement between unemployment and employment by looking at the spells of unemployment as a response variable and characteristics that one may relate to the length of spells. Some common characteristics that may be related to the unemployment spells are: age, sex, marital status of the individual, race (visible minority (VM) to be specific), education level, geographic location, health condition such as disable and/or partially handicap, and other socio-economic status such as individuals receiving employment insurance or social welfare.

When Milorad Kovacavic and myself were exploring the nature of the SLID data in last summer, we found that there are 35,669 individuals included in this longitudinal study. Many individuals, however, were found to belong to the groups with codes 96/6 or 97/7 or 98/8 or 99/9 recorded for some variables of interest. Here 96/6 (96 or 6; 2 or 1 digit codes for the variable concerned) refers to the individual 'not in sample,' 97/7 refers to the individual who responded 'don't know' to a question, 98/8 refers to the refusal, and similarly 99/9 refers to the individual with answer 'not applicable.' Consequently, for simplicity, we have decided to exclude all individuals with codes 96/6, 98/8 and 99/9 and to create 3 modified data files as follows. First set of modified data does not even contain the individuals with codes 97/7 for his or her answer, and there are 16,890 individuals in this group. This group will be referred to as the 'complete' group, both with respect to covariates

and the response variable. The second set of modified data contains the individuals of the first group plus those individuals with 97/7 code for the response variable, 'jobless spell,' and there are altogether 16,990 individuals under this group. Finally, the third set of modified data is created with individuals of the first group plus the individuals with 97/7 code either for covariates or response variable. We analyse the first set of data in Section 3 by using the methodologies developed in part I. In Section 4, we develop regression methodologies to deal with the second set of data where the outcomes are subject to nonresponse. The discussion about the third set of data where the outcomes and the covariates are subject to nonresponse is beyond the scope of the present report.

To shed some light on the nature of the longitudinal relationship between 'jobless spell' and the covariates, we first, under the assumption of equal weights, compute some basic statistics for the response variable for 1993 and 1994. These statistics are shown in Table 2.1

Table 2.1. Summary statistics for jobless spell (in weeks) for 1993 and 1994

Time	Minima	Maxima	Mean	Std Dev	Skewness	Kurtosis
1993	0	52	3.72	10.48	3.33	10.84
1994	0	52	3.93	10.96	3.22	9.89

It is clear from Table 2.1 that the jobless spells are highly positively skewed showing that while there are many individuals without or with minimum jobless spell, there also may be a considerable number of individuals with large jobless spell. To understand this more clearly, we now create 5 ordinal categories for jobless spell and exhibit the frequency distributions for 1993 and 1994 as in Table 2.2. The five ordinal categories are formed as follows. Let z_{it} denote the jobless period for the ith individual at the tth year. Thus z_{it} takes the values $z_{it} = 0, 1, 2, \ldots, 52$. Suppose we partition z_{it} values into the following 5 groups: $z_{it} = 0$; $1 \leq z_{it} \leq 4$; $5 \leq z_{it} \leq 12$; $13 \leq z_{it} \leq 26$; and $27 \leq z_{it} \leq 52$. Here, although the choice of cut points is arbitrary, it yields psychologically acceptable intervals, that is, 'less than a month', 'less than three months', 'less than half a year', and 'more than half a year'. Next suppose that $y_{ijt} = 1$ if z_{it} belongs to the jth ($j = 1, \ldots, 5$) group. This mechanism leads to the frequency distribution for the jobless spell as in Table 2.2.

Table 2.2. Percentage of individuals under 5 ordinal categories

	Ordinal categories for jobless spells					
	1	2	3	4	5	Total
1993	80.7	4.0	5.0	4.9	5.5	100
1994	80.8	3.9	4.5	4.9	6.0	100

As opposed to Table 2.1, Table 2.2 exhibits the longitudinal economic situation much more clearly. For example, it is clear from this table that although 81% individuals are employed during the whole year, 4 to 6 percentage of individuals appear to belong to each of the remaining 4 categories. When jobless spells of 1994 are compared with that of 1993, it is clear that there were 0.5% less individuals in category 3 in 1994 as compared to 1993. Conversely, in category 5, there appear 0.5% more individuals in 1994 as compared to 1993. Since category 5 refers to the unemployment group with no job for more than six months, this distribution in Table 2.2 indicates relatively worse economic situation in 1994 as compared to 1993.

Next, we explore to understand what may be the contributions of different covariates to form a frequency table for jobless spells as in Table 2.2, including the movement of the work force from one category to the other category during the 2 years of longitudinal period. For the purpose we have confined our analysis to six important covariates: age, sex, marital status, race (VM), education level and resident province of the individual. Next, appropriate cross tables are made for these covariates and the response variable, jobless spell, for 1993 and 1994. Table 2.3 shows the age and jobless spell relationship for 1993 and 1994. Table 2.4 shows the sex and jobless spell for 1993 and 1994, and so on.

The percentage of individuals are high in category 1 for all age groups, which is obvious. Under this category, however, more individuals in the 46-55 age group are full-time employed as compared to the other age groups.

Table 2.3. Cross-table for age group and jobless spell with cells indicating percentage of individuals

		Jobless spell category				
Age group	Year	1	2	3	4	5
20 or less	1993	7.80	0.98	1.17	0.37	0.25
	1994	5.67	0.94	1.10	0.49	0.38
21-25	1993	6.30	0.75	0.76	0.89	0.73
	1994	6.26	0.73	0.72	0.95	0.78
26-30	1993	9.28	0.50	0.61	0.81	0.83
	1994	8.53	0.49	0.56	0.76	0.79
31-35	1993	11.98	0.50	0.70	0.80	0.89
	1994	12.21	0.49	0.60	0.65	0.96
36-40	1993	11.71	0.38	0.62	0.70	0.88
	1994	12.21	0.49	0.54	0.60	0.89
41-45	1993	9.90	0.33	0.43	0.47	0.62
	1994	10.20	0.33	0.36	0.44	0.70
46-55	1993	14.79	0.42	0.45	0.54	0.91
	1994	15.53	0.32	0.49	0.72	0.99
56-65	1993	8.97	0.13	0.22	0.28	0.37
	1994	10.27	0.08	0.17	0.28	0.46
Total	1993	80.73	3.98	4.96	4.86	5.46
	1994	80.76	3.86	4.54	4.89	5.95

The percentage of individuals with year round jobs (category 1) is less in 1994 as compared to 1993 for the young age groups 25 or less. When umemployment distributions are compared category-wise, there are more individuals in category 5 than any other category in a given age group except for the youth groups up to age 25. These younger groups appear to have shorter jobless spell than the individuals in any other age group. When the jobless spells of 1994 are compared with those of 1993, there appear a decrease in percentage under category 3 but an increase in percentage under category 5, which is in agreement with Table 2.2. But it is clear from Table 2.3 that this pattern is not only valid for the total number of individuals under these two categories, it appears to hold uniformly across all age groups.

The observed marginal relationship for sex and the jobless spell for 1993 and 1994 are shown in Table 2.4. This table shows that a higher percentage of females is full-time employed as compared to males both for 1993 and 1994. As far as the longitudinal movement is concerned, the percentage

of males appear to decrease from 1993 to 1994 under categories 2, 3 and 4, and increase under category 5. This indicates a mixed economic situation. This is because although more males had employment in 1994, a large number of males also appear to belong in category 5 in 1994 as compared to 1993, which is a worse situation for this later group of males. A good percentage (0.25%) of females appeared to lose their full-time employment in 1994 as compared to 1993, and the percentage of females appear to increase from 1993 to 1994 under categories 2, 4, and 5, and decrease under category 3.

Table 2.4. Cross-table for sex and jobless spell with cells indicating percentage of individuals

		Jobless spell category				
Sex	Year	1	2	3	4	5
Male (=1)	1993	38.24	2.03	2.36	2.88	2.82
	1994	38.53	1.84	2.26	2.65	3.07
Female (=2)	1993	42.49	1.95	2.60	1.98	2.64
	1994	42.24	2.02	2.27	2.24	2.88

Next, the longitudinal relationship between marital status of an individual and the jobless spell for 1993 and 1994 are shown in Table 2.5. It is clear from Table 2.5 that a large percentage of married individuals are full-time employed followed by the single or never married group. There appear more longitudinal changes (when 94 compared with 93) in percentage of individuals for these two groups as compared to all other groups for all categories except the 5th category for married individuals. For the jobless spell part, a fewer percentage of individuals in a cell indicate better economic conditions. The single individuals who appeared to loose jobs in 1994 as compared to 1993 are seen to have larger jobless spells in 1994 as opposed to 1993. This leads to an overall mixed situation for these two larger groups (married and single groups). This is because more married group individuals appear to have full time jobs in 1994 as compared to 1993, wheras the single group individuals with year round jobs are less in 1994 as compared to 1993.

The observed marginal relationship for race and the jobless spell for 1993 and 1994 are reported in Table 2.6. Here by race, we mean the visible

Table 2.5. Cross-table for marital status and jobless spell with cells indicating percentage of individuals

		Jobless spell category				
Marital status	Year	1	2	3	4	5
Married (01)	1993	52.54	1.78	2.27	2.64	3.02
	1994	53.57	1.62	1.97	2.34	3.04
Common-law (02)	1993	4.87	0.26	0.36	0.49	0.54
	1994	4.96	0.25	0.37	0.47	0.62
Separated (03)	1993	1.91	0.07	0.09	0.16	0.21
	1994	2.14	0.12	0.15	0.22	0.27
Divorced (04)	1993	2.68	0.10	0.14	0.15	0.26
	1994	2.88	0.10	0.11	0.18	0.27
Widowed (05)	1993	1.30	0.02	0.05	0.05	0.05
	1994	1.46	0.05	0.04	0.05	0.08
Single (06)	1993	17.44	1.76	2.07	1.38	1.37
(never married)	1994	15.75	1.72	1.90	1.62	1.68
Total	1993	80.73	3.98	4.96	4.86	5.46
	1994	80.76	3.86	4.54	4.89	5.95

minority (VM) status of an individual, that is, the individual may or may not be a visible minority. Table 2.6 examines whether visible minority status affect the jobless spell or not, over the period.

Table 2.6. Cross-table for race and jobless spell with cells indicating percentage of individuals

		Jobless spell category				
VM status	Year	1	2	3	4	5
Yes (=1)	1993	2.66	0.12	0.24	0.12	0.23
	1994	2.71	0.14	0.14	0.17	0.22
No (=2)	1993	78.06	3.86	4.72	4.74	5.23
	1994	78.05	3.72	4.39	4.72	5.73

As the percentage of individuals in 1994 remains to be almost the same as in 1993 for any given jobless spell category, there does not appear any longitudinal movement either for the visible minority or majority group. However, the distribution pattern of individuals under different categories for the visible minority appears to be generally different than for the non-minority group.

Next in Table 2.7, we display the joint frequency distribution of the jobless spell and the highest education level of the individual. There are 12 levels of education beginning from 'never attended school' as level 1 to Doctorate (Ph.D.) as level 12.

Table 2.7 shows that for education levels from 1 to 3 and 7 to 12, there does not appear any changes in percentage of individuals in 1994 as compared to 1993. But, for the education level from 4 to 6, the proportion of individuals in 1994 are usually different than in 1993. This seems quite sensible as these levels of education refer to the individuals with 9 to 13 years of school as well as high school graduates. Thus, the jobless spells appear to be dependent on the education level of the individual. Our confirmatory analysis performed in the next section also supports this relationship between the jobless spell and the education level of the individual.

Table 2.7. Cross-table for education level and jobless spell with cells indicating percentage of individuals

		Jobless spell category				
Education level	Year	1	2	3	4	5
01	1993	0.08	0.01	0.00	0.01	0.00
	1994	0.09	0.01	0.00	0.00	0.00
02	1993	0.40	0.01	0.03	0.02	0.04
	1994	0.38	0.02	0.01	0.03	0.04
03	1993	5.28	0.32	0.29	0.44	0.72
	1994	5.38	0.17	0.24	0.41	0.75
04	1993	8.63	0.40	0.69	0.63	0.92
	1994	7.26	0.33	0.43	0.54	1.05
05	1993	6.20	0.32	0.62	0.33	0.37
	1994	5.42	0.32	0.55	0.33	0.45
06	1993	13.28	0.65	0.83	0.73	0.85
	1994	12.95	0.50	0.67	0.78	1.07
07	1993	6.49	0.52	0.44	0.58	0.48
	1994	6.74	0.49	0.60	0.71	0.53
08	1993	4.03	0.49	0.35	0.19	0.17
	1994	4.47	0.54	0.32	0.24	0.12
09	1993	24.30	7.00	1.31	1.58	1.55
	1994	25.46	1.02	1.34	1.50	1.60
10	1993	1.74	0.05	0.05	0.08	0.06
	1994	1.82	0.07	0.08	0.08	0.08
11	1993	6.79	0.18	0.26	0.21	0.20
	1994	7.08	0.34	0.24	0.18	0.20
12	1993	3.51	0.05	0.08	0.06	0.11
	1994	3.71	0.06	0.07	0.08	0.07
Total	1993	80.73	3.98	4.96	4.86	5.46
	1994	80.76	3.86	4.54	4.89	5.95

Table 2.8. Cross-table for province of residence and jobless spell with cells indicating percentage of individuals

		Jobless spell category				
Province	Year	1	2	3	4	5
Newfoundland (10)	1993	4.49	0.40	0.47	0.44	0.94
	1994	4.38	0.37	0.41	0.46	1.03
Prince Edward	1993	1.71	0.14	0.10	0.18	0.14
Island (11)	1994	1.66	0.07	0.10	0.21	0.19
Nova Scotia (12)	1993	5.16	0.33	0.22	0.34	0.33
	1994	4.99	0.25	0.30	0.36	0.41
New Brunswick (13)	1993	4.57	0.27	0.34	0.34	0.36
	1994	4.62	0.22	0.31	0.28	0.44
Quebec (24)	1993	16.96	0.74	1.18	1.20	1.43
	1994	16.76	0.82	0.97	1.29	1.69
Ontario (35)	1993	20.62	0.83	1.10	1.04	1.20
	1994	20.89	0.80	0.99	1.11	1.10
Manitoba (46)	1993	5.67	0.27	0.36	0.25	0.23
	1994	5.61	0.35	0.33	0.24	0.20
Saskatchewan (47)	1993	6.29	0.21	0.31	0.26	0.24
	1994	6.26	0.25	0.29	0.25	0.21
Alberta (48)	1993	8.14	0.42	0.37	0.45	0.34
	1994	8.20	0.38	0.40	0.35	0.37
British Columbia (59)	1993	7.09	0.36	0.47	0.34	0.27
	1994	7.32	0.32	0.41	0.32	0.31
Other (60)	1993	0.05	0.01	0.04	0.01	0.01
	1994	0.08	0.02	0.02	0.01	0.01
Total	1993	80.73	3.98	4.96	4.86	5.46
	1994	80.76	3.86	4.54	4.89	5.95

Finally, the relationship between geographical location of the individual and the jobless spell for 1993 and 1994 are displayed in Table 2.8. This table shows that except for Ontario, Quebec and British Columbia, the percentage of individuals in 1994 for a given category of jobless spell remain almost the same as in 1993. Between Ontario and Quebec, the percentage of individuals appear to increase in Quebec for 1994 as compared to 1993, for almost all categories from 2 to 5, whether for Ontario, the percentage of individuals appear to decrease. The percentage of full-time employments (category 1) appears to be larger in 1994 as compared to 1993 for Ontario but the percentage of year round jobs goes down for Quebec. The jobless spell situation in British Columbia appears to

be similar to that in Ontario.

The purpose of the next section is to examine the joint behaviour of the six covariates: age, sex, marital status, race (VM), education level, and province of residence in explaining the distribution of the jobless spell. This is done by applying the regression methodology developed in part I.

3 Regression Analysis For SLID Data

In this section we analyze the complete SLID data in two parts. First, we examine the effect of the covariates on jobless spells based on complete information for a given year, say 1993. This is referred to as the cross-sectional analysis, although we have only information on one individual from a household (as opposed to information from all co-habitants of the unit) who is selected as the longitudinal unit for the complete period during the panel. This analysis, alternatively, may be referred to as the marginal regression analysis, where the data for a given year are the marginal data when compared with a whole longitudinal data set collected over the duration of the panel. Second, we deal with the longitudinal data for two years 1993 and 1994, which is a special case of the general multivariate regression analysis developed in part I.

3.1 Cross-sectional analysis with and without survey weights

Here we compute the regression effects of the covariates on the jobless spell for 1993 and 1994 separately. In notation of part I, let β_j be the p-dimensional regression parameter vector, where $j = 1, \ldots, J$, J being the number of categories for jobless spell. As we are considering 6 covariates: age (X_1), sex (X_2), marital status (X_3), race (VM) (X_4), education level (X_5) and province of residence (X_6), we have $p = 6$. Further, as the jobless spells were partitioned into 5 ordinal categories, we have $J = 5$. For a given year the estimating equations (cf. eqn (4.2) in part I) for the regression vector may be written as

$$\sum_{i=1}^{I} w_{its^*} u_{it}^*(\beta) = 0, \tag{3.1}$$

where $I = 16890$ and w_{its^*} refers to the cross-sectional survey weight for the ith longitudinal unit selected for sample s^* at a given year t (1993 or 1994), and $u_{it}^*(\beta)$ is given by

$$u_{it}^*(\beta) = W_{it}^{*T} \Sigma_{it}^{*-1} (Y_{it}^* - \mu_{it}^*(\beta)), \tag{3.2}$$

where

$$\begin{aligned} Y_{it}^* &= (y_{it1}, \ldots, y_{itj}, \ldots, y_{it,J-1})^T, \mu_{it}^*(\beta) = (\mu_{it1}, \ldots, \mu_{itj}, \ldots, \mu_{it,J-1})^T \\ W_{it}^{*T} &= \partial (Y_{it}^* - \mu_{it}^*(\beta))^T / \partial \beta, \end{aligned}$$

and Σ_{it}^* is the $(J-1)\times(J-1)$ covariance matrix of Y_{it}^*. More specifically, in (3.2), for $j = 1, \ldots, J-1$,

$$\mu_{itj} = \exp\{x_{it}^* \beta_j\} / \sum_{r=1}^{J} \exp\{x_{it}^T \beta_r\} \tag{3.3}$$

with $\beta_J = \beta_5 = 0$ without any loss of generality, and Σ_{it}^* is given by

$$\Sigma_{it}^* = \begin{bmatrix} \mu_{it1}(1-\mu_{it1}) & -\mu_{it1}\mu_{it2}, & \cdots, & -\mu_{it1}\mu_{it,J-1} \\ -\mu_{it2}\mu_{it1} & \mu_{it2}(1-\mu_{it2}), & \cdots, & -\mu_{it2}\mu_{it,J-1} \\ \vdots & \vdots & & \vdots \\ -\mu_{it,J-1}\mu_{it1} & -\mu_{it,J-1}\mu_{it2} & \cdots & \mu_{it,J-1}(1-\mu_{it,J-1}) \end{bmatrix},$$

which is the covariance matrix for the logistic multinomial distribution with J responses for the ith individual. Further, in (3.2), W_{it}^{*T} is simplified as

$$W_{it}^{*T} = \Sigma_{it}^* \otimes x_{it} \tag{3.4}$$

where $\otimes$ denotes the Kronecker product and $x_{it} = (X_{it1}, \ldots, X_{itu}, \ldots, X_{itp})^T$ is the $p \times 1$ ($p = 6$) vector of the covariates for the ith individual at the tth time.

The solution of (3.1), denoted by $\hat{\beta} = (\hat{\beta}_1^T, \ldots, \hat{\beta}_j^T, \ldots, \hat{\beta}_{J-1}^T)^T$, is obtained iteratively by using the iterative equation

$$\begin{aligned} \hat{\beta}(m+1) &= \hat{\beta}(m) + \left[\sum_{i \epsilon s^*} w_{its^*} W_{it}^{*T} \Sigma_{it}^{*-1} W_{it}^* \right]_m^{-1} \\ &\quad \times \left[\sum_{i \epsilon s^*} w_{its^*} W_{it}^{*T} \Sigma_{it}^{*-1} (Y_{it}^* - \mu_{it}^*(\beta)) \right]_m, \end{aligned} \tag{3.5}$$

where $[\cdot]_m$ denotes that the expression within the brackets is evaluated at $\hat{\beta}(m)$, the value of $\hat{\beta}$ at the mth iteration. Under some mild conditions, $\hat{\beta}$ is asymptotically normal with mean β and covariance matrix $\text{cov}(\hat{\beta})$ which can be consistently estimated by

$$\hat{\text{cov}}(\hat{\beta}) = \left[\sum_{i\epsilon s^*} w_{its^*} W_{it}^{*T} \hat{\Sigma}_{it}^{*-1} W_{it}^*\right]^{-1}, \tag{3.6}$$

where Σ_{it}^* is given in (3.2). Note that the estimate of $\text{cov}(\hat{\beta})$ in (3.6) is obtained by observing the fact that under census equation, one computes the covariance matrix of the well-known quasi-likelihood estimator of $\hat{\beta}$ as

$$\begin{aligned}
\text{cov}(\hat{\beta}) &= \left[\sum_{i=1}^{I} W_{it}^{*T} \Sigma_{it}^{*-1} W_{it}^*\right]^{-1} \\
&\quad \times \left[\sum_{i=1}^{I} W_{it}^{*T} \Sigma_{it}^{*-1} E(Y_{it}^* - \mu_{it}^*(\beta))(Y_{it}^* - \mu_{it}^*(\beta))^T \Sigma_{it}^{*-1} W_{it}^*\right] \\
&\quad \times \left[\sum_{i=1}^{I} W_{it}^{*T} \Sigma_{it}^{*-1} W_{it}^*\right]^{-1} \\
&= \left[\sum_{i=1}^{I} W_{it}^{*T} \Sigma_{it}^{*-1} W_{it}^*\right]^{-1} \left[\sum_{i=1}^{I} W_{it}^{*T} \Sigma_{it}^{*-1} W_{it}^*\right] \\
&\quad \times \left[\sum_{i=1}^{I} W_{it}^{*T} \Sigma_{it}^{*-1} W_{it}\right]^{-1}.
\end{aligned} \tag{3.7}$$

Note that for equal survey weights or $w_{its^*} = 1$ (say), the estimate of the covariance matrix in (3.6) reduces to

$$\hat{\text{cov}}(\hat{\beta}) = \left[\sum_{i=1}^{I} W_{it}^{*T} \hat{\Sigma}_{it}^{*-1} W_{it}^*\right]^{-1}. \tag{3.8}$$

Now by using the iterative equation (3.5) and covariance matrix estimate in (3.6), we obtain the regression estimates for 6 covariates and their standard errors for the case with survey weights. Next by using $w_{its^*} = 1$ in (3.5) and the covariance matrix estimate in (3.8), we obtain the regression estimates for 6 covariates and their standard errors for the case without the survey weights. Both of

these regression estimates, that is, estimates based on with and without survey weights, along with their standard errors are reported in Table 3.1.

Further note that while using survey weights in (3.5) and (3.6), it is assumed that one only knows the survey weights and there is no information available about the complex design that has been used to generate these survey weights. For the case when complex survey design is known to be stratified cluster sampling, the regression estimates still are computed by using (3.5) but their covariance matrix is now estimated by

$$\begin{aligned}
\hat{\text{cov}}(\hat{\beta}) &= \left[\sum_{i\epsilon s^*} w_{its^*} W_{it}^{*T} \Sigma_{it}^{*-1} W_{it}^*\right]^{-1} \text{cov}\left[\sum_{i\epsilon s^*} w_{its^*} W_{it}^{*T} \Sigma_{it}^{*-1} (Y_{it}^* - \mu_{it}^*(\beta))\right] \\
&\quad \times \left[\sum_{i\epsilon s^*} w_{its^*} W_{it}^{*T} \Sigma_{it}^{*-1} W_{it}^*\right]^{-1} \\
&= \left[\sum_{i\epsilon s^*} w_{its^*} W_{it}^{*T} \Sigma_{it}^{*-1} W_{it}^*\right]^{-1} \left[\sum_{h=1}^{L} \{n_h(1-n_h)^{-1}\} \sum_{c=1}^{n_h} (z_{hct} - \bar{z}_{ht})(z_{hct} - \bar{z}_{ht})^T\right] \\
&\quad \times \left[\sum_{i\epsilon s^*} w_{its^*} W_{it}^{*T} \Sigma_{it}^{*-1} W_{it}\right]^{-1}, \qquad (3.9)
\end{aligned}$$

where $w_{its^*} = w_{hcits^*}$ and $z_{its^*} = W_{it}^{*T} \Sigma_{it}^{*-1} (Y_{it} - \mu_{it}(\beta)) = z_{hcits^*}$, when information on the h-th stratum and the c-th cluster is available, and

$$z_{hct} = \sum_{i=1}^{n_{hc}} w_{hcits^*} z_{hcits^*}, \text{ and } \bar{z}_{ht} = \sum_{c=1}^{n_h} z_{hct}/n_h.$$

In the present analysis we have used the survey weights only as there were no information available about the complex survey design.

Table 3.1. Cross-sectional regression estimates (upper entry) with multinomial logistic margins and their standard errors (lower entry) with or without survey weights (SW)

			Covariates					
Year	SW	Jobless spell	X_1	X_2	X_3	X_4	X_5	X_6
1993	Without	1	-0.027	0.206	-1.044	-8.112	0.921	0.395
			0.003	0.065	0.018	0.096	0.013	0.002
		2	-0.138	-0.087	-0.869	-6.712	0.746	0.370
			0.004	0.099	0.003	0.144	0.020	0.003
		3	-0.140	0.474	-0.592	-6.744	0.676	0.381
			0.004	0.101	0.027	0.148	0.020	0.003
		4	-0.099	-0.908	-0.987	-6.345	0.741	0.360
			0.004	0.088	0.024	0.129	0.017	0.003
	With	1	-0.099	0.080	-1.292	-5.867	0.976	0.304
			0.006	0.121	0.033	0.180	0.024	0.004
		2	-0.165	-0.188	-1.184	-5.587	0.849	0.300
			0.008	0.166	0.046	0.247	0.033	0.006
		3	-0.161	-0.059	-1.118	-5.890	0.897	0.300
			0.007	0.156	0.043	0.231	0.031	0.005
		4	-0.148	-0.515	-1.245	-5.358	0.854	0.290
			0.007	0.152	0.042	0.225	0.030	0.005
1994	Without	1	-0.010	-0.179	-1.573	-11.083	1.341	0.441
			0.003	0.068	0.019	0.101	0.014	0.002
		2	-0.077	-0.172	-1.332	-10.810	1.252	0.426
			0.004	0.099	0.027	0.145	0.020	0.003
		3	-0.077	-0.271	-1.325	-10.375	1.190	0.425
			0.004	0.099	0.027	0.145	0.020	0.003
		4	-0.052	-0.510	-1.360	-10.269	1.174	0.409
			0.004	0.096	0.026	0.141	0.019	0.003
	With	1	-0.174	-3.108	-2.357	-5.036	1.345	0.355
			0.014	0.306	0.080	0.404	0.062	0.012
		2	-0.281	-2.641	-1.735	-5.842	1.288	0.381
			0.019	0.416	0.109	0.550	0.084	0.016
		3	-0.269	-3.675	-1.925	-4.087	1.215	0.339
			0.016	0.359	0.094	0.475	0.073	0.014
		4	-0.266	-3.758	-1.680	-2.683	0.877	0.356
			0.018	0.402	0.105	0.532	0.082	0.016

It is clear from Table 3.1 that the regression estimates for the age group (X_1) are negative with small standard errors. Thus, these estimates appear to be significant under all categories both for 1993 and 1994. As the marginal probabilities are in the exponential (logistic) form, a negative regression estimate for X_1 indicates that a higher age group individual has less chance to be in the

category under consideration as compared to a younger age group. When the regression estimates for X_1 are compared across four categories, they appear to have smaller negative values under category 1 as compared to other categories. This shows that the probability for any individual to be in category 1 is higher than his or her probability to be in any other category. This is obvious and it agrees with Table 2.3 as there are more full-time employed individuals (in category 1) as compared to jobless or unemployed individuals. When jobless spells are compared, an individual has almost equal chances to be in any of the 3 other categories from category 2 to category 4. This means that if an individual is unemployed, it is almost equally likely that s/he may be without a job for less than 4 or 12 or 26 weeks, with a slightly higher chance being in the longer jobless spell group of up to 26 weeks. This appears to hold for both 1993 and 1994, the situation in 1994 is being slightly worse as compared to 1993. This is because in 1994, the regression estimates due to the age of individuals are large negative as compared to 1993. When regression estimates computed with or without survey weights are compared, the regression methodology without using survey weights appears to produce under estimates for the standard errors of regression effects which indicate that the introduction of the survey weights perhaps decreases the bias of the estimate. A simulation study to this effect could be done to see the extent of bias reduction due to the use of survey weights but such studies are not chosen in the present report.

To examine the effects of sex on the jobless spell, consider the regression estimates for category 1 computed based on survey weights. A small positive estimate (0.080) in 1993 indicates a larger probability for a female (code 2) to be included in the full-time job category as compared to a male (with code 1). In 1994, however, the situation gets reversed as the regression estimate is large negative (-3.108). This implies that in 1994, a male had larger probability than a female to be included into the full-time employment category. As far as the jobless durations are concerned, females appear to have smaller probabilities to be in any of the remaining 3 categories, which is true for both 1993 and 1994 in general. The females those who moved from full-time job category in 1993 to other jobless spell categories in 1994, are understood by comparing the regression estimates in

category 1 with other regression estimates in other categories both in 1993 and 1994. For example, while a female had more probability to be included in category 1 in 1993 (with regression estimate 0.080), in 1994, she has larger probability to be included in category 2 (with regression estimate -2.641) as compared to category 1 (with regression estimate -3.108). Similarly, one may explain the employment movement of a male during the 1993-94 period.

All negative estimates for the marital status (X_3) and visible minority (X_4) covariates appear to suggest that as the levels (codes) are increasing for these covariates, the probabilities will get smaller for the individuals to get included in a given category. Furthermore, as the estimates do not appear to change significantly from 1993 to 1994, they do not appear to have any substantial impact on the movement of the work force.

Finally, for the last two covariates X_5 (education level) and X_6 (province of residence status), the regression estimates are positive. These estimates are larger in magnitude in 1994 as compared to 1993. Thus, for a higher educated individual, the probability of the individual to be in category 1 (year round employment) is higher in 1994 than 1993. Although s/he has chance to be included in the other categories, the probability for this to happen is smaller as the regression values for categories 2 through 4 are smaller as compared to the regression value under category 1. The effect of the provincial residence status may similarly be explained.

In the next section, we combine the information from 1993 and 1994 surveys and examine the effects of covariates on the jobless spell.

3.2 Longitudinal analysis with and without survey weights

Let $Y_i = [Y_{i1}^{*T}, Y_{i2}^{*T}]^T$ be the vector of ordinal responses for the ith person for 1993 and 1994. Here $Y_{i1}^* = (y_{i11}, \ldots, y_{i1j}, \ldots, y_{i1,J-1})^T$ refers to the multinomial response for jobless spell for the year 1993 and similarly Y_{i2}^* is defined for 1994. Since $J = 5$, Y_i is a 8-dimensional response vector. Let μ_i be the mean vector of Y_i so that $\mu_i = [\mu_{i1}^{*T}, \mu_{i2}^{*T}]^T$, where for $t = 1, 2$, $\mu_{it}^* = (\mu_{it1}, \ldots, \mu_{itj}, \ldots, \mu_{it,J-1})^T$ is given by (3.2), with μ_{itj} as defined in (3.3).

As Y_{i1}^* and Y_{i2}^* are two multinomial response vector for the same ith individual, it is likely that they will be correlated. Suppose that this longitudinal correlation is denoted by α_{12} irrespective of the category of the responses. Then the covariance matrix of Y_i may be written as

$$\Sigma_i(\alpha_{12}) = \begin{bmatrix} \Sigma_{i1}^* & \alpha_{12}\Sigma_{i12}^* \\ \alpha_{12}\Sigma_{i12}^{*T} & \Sigma_{i2}^* \end{bmatrix}, \tag{3.10}$$

where for $t = 1, 2$, Σ_{it}^* is the covariance matrix for the responses of the ith individual given at tth year. Since $\alpha_{12}U_{J-1}$ is the correlation matrix for Y_{i1}^* and Y_{i2}^*, where U_{J-1} is the $(J-1)\times(J-1)$ unit matrix, Σ_{i12}^* may then be written as

$$\Sigma_{i12}^* = \begin{bmatrix} \lambda_{i11}^{\frac{1}{2}}\lambda_{i21}^{\frac{1}{2}} & \lambda_{i11}^{\frac{1}{2}}\lambda_{i22}^{\frac{1}{2}} & \cdots & \lambda_{i11}^{\frac{1}{2}}\lambda_{i2,J-1}^{\frac{1}{2}} \\ \lambda_{i12}^{\frac{1}{2}}\lambda_{i21}^{\frac{1}{2}} & \lambda_{i12}^{\frac{1}{2}}\lambda_{i22}^{\frac{1}{2}} & \cdots & \lambda_{i12}^{\frac{1}{2}}\lambda_{i2,J-1}^{\frac{1}{2}} \\ \vdots & \vdots & & \vdots \\ \lambda_{i1,J-1}^{\frac{1}{2}}\lambda_{i21}^{\frac{1}{2}} & \lambda_{i1,J-1}^{\frac{1}{2}}\lambda_{i22}^{\frac{1}{2}} & \cdots & \lambda_{i1,J-1}^{\frac{1}{2}}\lambda_{i2,J-1}^{\frac{1}{2}} \end{bmatrix}, \tag{3.11}$$

where for $t = 1, 2$, $\lambda_{itj} = \mu_{itj}(1-\mu_{itj})$ for all $j = 1, \ldots, J-1$, where μ_{itj} is the mean of Y_{itj}.

For known α_{12}, the regression effects of the six covariates may now be computed by solving the estimating equations

$$\sum_{i=1}^{I} w_{i\ell s^*} u_{it}(\beta) = 0, \tag{3.12}$$

where $w_{i\ell s^*}$ is the longitudinal weight for the ith individual included in the sample s^*, and $u_{it}(\beta)$ is given by

$$u_{it}(\beta) = W_i^T \Sigma_i^{-1}(\alpha_{12})(Y_i - \mu_i), \tag{3.13}$$

which is quite similar to (3.2). Here

$$W_i^T = [W_{i1}^{*T}, W_{i2}^{*T}],$$

where

$$W_{i1}^{*T} = \partial(Y_{i1}^* - \mu_{it})/\partial\beta^T, \text{ and } W_{i2}^{*T} = \partial(Y_{i2}^* - \mu_{i2})/\partial\beta^T.$$

Consequently, in the manner similar to that of (3.5), the regression effects are computed by using the iterative equation

$$\beta^*(m+1) = \beta^*(m) + \left[\sum_{i\epsilon s^*} w_{i\ell s^*} W_i^T \Sigma_i^{-1}(\alpha_{12}) W_i\right]^{-1} \times \left[\sum_{i\epsilon s^*} w_{i\ell s^*} W_i^T \Sigma_i^{-1}(\alpha_{12})(Y_i - \mu_i)\right]_m . \quad (3.14)$$

where $[\cdot]_m$ denotes that the expression within the brackets is evaluated at $\beta^*(m)$, the value of β^* at the mth iteration.

Next, following (3.6), the covariance matrix of β^* may be consistently estimated by

$$\hat{\text{cov}}(\beta^*) = \left[\sum_{i\epsilon s^*} w_{i\ell s^*} W_i^T \Sigma_i^{-1}(\alpha_{12}) W_i\right]^{-1} \quad (3.15)$$

For equal survey weights, that is, when $w_{i\ell s^*} = 1$, the covariance matrix estimate (3.14) reduces to

$$\hat{\text{cov}}(\beta^*) = \left[\sum_{i=1}^{I} W_i^T \Sigma_i^{-1}(\alpha_{12}) W_i\right]^{-1} . \quad (3.16)$$

Note that in practice α_{12} is unknown. Following Sutradhar and Das (1999), we may estimate this longitudinal correlation by

$$\hat{\alpha}_{12} = \frac{\sum_{i\epsilon s^*} \sum_{j=1}^{J-1} w_{i\ell s^*}(y_{i1j} - \hat{\mu}_{i1j})(y_{i2j} - \hat{\mu}_{i2j})}{\left\{\sum_{i\epsilon s^*} \sum_{j=1}^{J-1} w_{i\ell s^*}(y_{i1j} - \hat{\mu}_{i1j})^2\right\}^{\frac{1}{2}} \left\{\sum_{i\epsilon s^*} \sum_{j=1}^{J-1} w_{i\ell s^*}(y_{i2j} - \hat{\mu}_{i2j})^2\right\}^{\frac{1}{2}}}, \quad (3.17)$$

which is consistent for α_{12}.

For the case when complex survey design is known to be stratified cluster sampling, the regression estimates still are computed by using (3.14) and their covariance matrix may be estimated as in the cross-sectional set up (cf. eqn. 3.9). The only difference is that now we compute z_{hc} in place of z_{hct} based on the longitudinal weights as follows. Here $w_{i\ell s^*} = w_{hci\ell s^*}$, and for known α_{12}, $z_{i\ell s^*} = W_i^{*T} \Sigma_i^{*-1}(\alpha_{12})(Y_i - \mu_i(\beta)) = z_{hci\ell s^*}$, when information on the h-th stratum and the c-th cluster is available, and

$$z_{hc} = \sum_{i=1}^{n_{hc}} w_{hci\ell s^*} z_{hci\ell s^*}, \text{ and } \bar{z}_h = \sum_{c=1}^{n_h} z_{hc}/n_h.$$

Note, however, that since α_{12} is unknown in practice, one requires to estimate this, based on the complex survey design. When a stratified cluster sampling design is used, this longitudinal correlation may be estimated as

$$\hat{\alpha}_{12} = \frac{\sum_{h=1}^{L}\sum_{c=1}^{n_h}(n_h/(n_h-1))(y_{hc1}-\bar{y}_{h1})(y_{hc2}-\bar{y}_{h2})}{\left\{\sum_{h=1}^{L}\sum_{c=1}^{n_h}(n_h/(n_h-1))(y_{hc1}-\bar{y}_{h1})^2\right\}^{\frac{1}{2}}\left\{\sum_{h=1}^{L}\sum_{c=1}^{n_h}(n_h/(n_h-1))(y_{hc2}-\bar{y}_{h2})^2\right\}^{\frac{1}{2}}}, \tag{3.18}$$

where, for $t = 1$, $y_{hc1} = \sum_{i=1}^{n_{hc}} w_{hci\ell s^\bullet} y_{hci\ell s^\bullet}$, and $\bar{y}_{h1} = \sum_{c=1}^{n_h} y_{hc1}/n_h$. Similarly, for $t = 2$, one defines y_{hc2} and $\bar{y}_{h2}$.

In the present analysis we have used the survey weights only as there were no information available about the complex survey design. Thus, the applications of (3.14), (3.15) and (3.16) to the SLID data for 1993 and 1994 yield the regression estimates and the estimates of their standard errors as in Table 3.2. The longitudinal analysis reported in Table 3.2 reveals an interesting finding that the regression effects of the covariates and their standard error estimates are almost the same for categories 2, 3, and 4. other.

Table 3.2. Longitudinal regression estimates (upper entry) with multinomial logistic margins and their standard errors (lower entry) with or without survey weights (SW)

			Covariates					
Approach	SW	Jobless spell	X_1	X_2	X_3	X_4	X_5	X_6
Independence	With	1	-0.060	1.266	-1.762	-12.707	1.696	0.448
($\alpha_{12} = 0$)			0.008	0.175	0.046	0.230	0.035	0.007
		2	-0.093	1.030	-1.736	-13.159	1.624	0.443
			0.010	0.225	0.059	0.297	0.045	0.009
		3	-0.093	1.029	-1.728	-13.173	1.627	0.442
			0.010	0.224	0.058	0.295	0.045	0.009
		4	-0.092	0.988	-1.726	-13.085	1.613	0.442
			0.010	0.225	0.058	0.296	0.045	0.009
	Without	1	-0.041	0.443	-1.676	-15.702	1.806	0.582
			0.001	0.032	0.009	0.047	0.006	0.001
		2	-0.067	0.302	-1.628	-16.256	1.734	0.570
			0.003	0.064	0.017	0.095	0.013	0.002
		3	-0.067	0.313	-1.625	-16.233	1.729	0.571
			0.003	0.065	0.018	0.095	0.014	0.002
		4	-0.065	0.249	-1.639	-16.183	1.729	0.569
			0.003	0.063	0.017	0.093	0.013	0.002
Generalized	With	1	-0.065	1.282	-1.805	-12.751	1.784	0.440
($\hat{\alpha}_{12} = 0.763$)			0.078	1.723	0.458	2.284	0.353	0.068
		2	-0.097	1.044	-1.779	-13.236	1.714	0.434
			0.084	1.862	0.494	2.464	0.380	0.073
		3	-0.097	1.044	-1.771	-13.252	1.716	0.434
			0.084	1.862	0.494	2.464	0.381	0.073
		4	-0.095	1.007	-1.769	-13.154	1.703	0.434
			0.084	1.858	0.492	2.459	0.380	0.079
($\hat{\alpha}_{12} = 0.733$)	Without	1	-0.047	0.464	-1.742	-15.455	1.799	0.579
			0.005	0.102	0.028	0.152	0.021	0.003
		2	-0.073	0.335	-1.694	-15.983	1.726	0.568
			0.007	0.159	0.044	0.236	0.032	0.005
		3	-0.074	0.338	-1.688	-15.960	1.720	0.568
			0.007	0.160	0.044	0.237	0.032	0.005
		4	-0.071	0.276	-1.701	-15.891	1.721	0.566
			0.007	0.156	0.043	0.232	0.032	0.005

This shows that it may be sufficient to partition the jobless spells into two categories only, namely, employed over the year and jobless for sometimes upto a year. This result appears to hold under both independence and generalized estimating equations approaches, irrespective of the use

of survey weights in the estimation of the regression parameters. As in the cross-sectional analyses, the use of survey weights changes the estimates of the standard errors.

By using the generalized estimating equation approach we find $\hat{\alpha}_{12} = 0.763$ when survey weights are used, and $\hat{\alpha}_{12} = 0.733$ when survey weights are not used. We compare the regression effects of the covariates yielded by the independence and generalized approaches, when survey weights are used. It appears from the first and third blocks of Table 3.2 that in the independence approach, all 6 covariates appear significant to explain the jobless spells, but the generalized approach shows that age (X_1) and sex (X_2) effects are not different from zero. The latter follows from the fact that the estimates of the standard errors are quite large as compared to the actual regression estimates for X_1 and X_2 under the generalized approach. For example, for X_1, for category 1 the regression estimate is -0.065 and its standard error is 0.078 leading to the conclusion that the age is an insignificant covariate to explain the jobless spell. This perhaps can be explained from the exploratory Table 2.1 when it was seen that the workforce movements were not very different among the higher and lower age groups.

With regard to the other 4 covariates, the regression effects for both category 1 and any other category appear to be negative for X_3 (marital status) and X_4 (race, VM), and positive for X_5 (education level) and X_6 (province of residence). Thus the marginal probabilities due to X_5 and X_6 will be considerably high and will be considerably low due to X_3 and X_4. Thus the education level and province of residence appear to be the dominant covariates for the movement of workforce over this small duration (1993 and 1994) of the survey.

4 Further Regression Analysis With Outcomes Subject to Nonresponse

In longitudinal studies, outcomes that are repeatedly measured over time may be correlated and some may be missing. The generalized estimating equations approach applied in the previous section to analyse the ordinal longitudinal data assumes that there are no missing data. In this

section, we deal with the situations when some of the longitudinal outcomes may be missing. To do this, it is, however, important to know the nature of the missingness of the data. First, the data may be missing completely at random (MCAR), which means that missingness is not related to the variables under study. Second, they may be missing at random (MAR), which means that missingness is related to the observed data but not to the missing data. Finally the nature of the missingness may be nonignorable, which means that missingvalues can not be generally ignored. There are some studies that deal with missing data problems in longitudinal analysis. For example, we refer to Kenward, Lesaffre and Molenburgs (1994), Robins, Rotnitzky and Zhao (1995), and Paik (1997). These authors have extended the idea of using the estimating equations approach of Liang and Zeger (1986) to the longitudinal data analysis in the presence of missing data. In the following sections, we suggest some modifications to their approaches, in particular to the recent approach of Paik (1997) in order to handle the missing data problem in the context of SLID data analysis through regression methods.

4.1 Generalized estimating equations (GEE) when data are missing completely at random (MCAR)

To initiate the discussion on the missing data problems, let $Y_i^c = (Y_{i1}^{*T}, \ldots, Y_{it_i}^{*T}, \ldots, Y_{iT_0}^{*T})^T$ and $x_i^c = (x_{i1}, \ldots, x_{it_i}, \ldots, x_{iT_0})^T$ denote the complete outcomes and covariates for the ith $(i = 1, \ldots, I)$ individual. Notice that we are now using a slightly different notation than in part I or the last sections of the sequel. To be specific, it will be assumed in the missing data case that t_i longitudinal outcomes will be observed for the i individual, and other outcomes (i.e. $T_0 - t_i$) will be missing, where as in part I, T_0 is the duration of the longitudinal survey. Further, for $t = 1, \ldots, T_0$, $r_{it} = 1$ if Y_{it}^* is observed, and $r_{it} = 0$ otherwise. Then the estimating equations in (3.10) can be re-expressed as

$$\sum_{i \epsilon s^*} w_{i\ell s^*} W_i^{c^T} \Sigma_i^{c^{-1}}(\alpha) R_i (Y_i^c - \mu_i^c) = 0, \tag{4.1}$$

where $R_i = I_{J-1} \otimes \text{diag}\{r_{it}\}$ with order $(J-1)T_0$, and $\Sigma_0^c(\alpha) = \text{var}(Y_i^c|x_i^c)$. If the data are missing completely at random (MCAR) then

$$E(Y_{itj}|x_{it}, r_{it} = 1) = E(Y_{itj}|x_{it}, r_{it} = 0) = \mu_{itj},$$

for all categories $j = 1, \ldots, J-1$, and the inference conditional on R_i i.e. on $\{r_{it}\}$ is valid, because the estimating function has mean zero. Now, as mentioned earlier, if the ith individual has $r_{it} = 1$ for $t = 1, \ldots, t_i$ and $r_{it} = 0$ for $t = t_i + 1, \ldots, T_0$, then the estimating equations (4.1) reduce to

$$\sum_{i \epsilon s^*} w_{i\ell s^*} W_{it_i}^T \Sigma_{it_i}^{-1}(\alpha)(Y_{it_i}^* - \mu_{it_i}) = 0, \tag{4.2}$$

which is truly an unbalanced form of the estimating equations (3.12). In such unbalanced cases, the estimation of the longitudinal correlation parameters (denoted by α) is, however, not easy. Paik (1997) has mainly dealt with the MAR and non-ignorable cases with $\alpha = 0$ (working independence approach) and thus avoided the estimation of the longitudinal correlation parameters.

In the MCAR case, we suggest to estimate the correlation parameters as in the following. As T_0 is usually small, for example, $T_0 = 2$ for the SLID data discussed in Section 2. Let $\alpha_{tt'}$ be the longitudinal correlation for the responses collected at time points t and t'. Here $t \neq t'$ and $t, t' = 1, 2, \ldots, T_0$. If all I individuals were present at all time points, then $\alpha_{tt'}$ could be consistently estimated by

$$\hat{\alpha}_{tt'} = \frac{\sum_{i \epsilon s^*}^{I} \sum_{j=1}^{J-1} w_{i\ell s^*}(y_{itj} - \hat{\mu}_{itj})(y_{it'j} - \hat{\mu}_{it'j})}{\left\{\sum_{i \epsilon s^*}^{I} \sum_{j=1}^{J-1} w_{i\ell s^*}(y_{itj} - \hat{\mu}_{itj})^2\right\}^{\frac{1}{2}} \left\{\sum_{i \epsilon s^*}^{I} \sum_{j=1}^{J-1} w_{i\ell s^*}(y_{it'j} - \hat{\mu}_{it'j})^2\right\}^{\frac{1}{2}}}, \tag{4.3}$$

which is quite similar to (3.17). Now, in the MCAR case, as the ith individual has the information for t_i time points, the construction of $\Sigma_{it_i}(\alpha)$ requires the estimation of $\alpha_{tt'}$ for $t \neq t'$ with $t, t' = 1, \ldots, t_i$. Here $t_i \leq T_0$. In order to obtain the correlation matrix for the ith individual, we compute the complete correlation matrix $\alpha_{tt'}$ based on the responses from $I_{tt'}$ individuals. Here $I_{tt'} \leq I$ is the number of individuals who provide the information both at time points t and t'. Altogether, there are $T_0(T_0 - 1)/2$ distinct correlation parameters to be estimated. Thus for $T_0 = 2, 3$ and 4, for

example, the number of parameters are 1, 3 and 6 respectively. Note that $\alpha_{tt'}$ computed by

$$\alpha^*_{tt'} = \frac{\sum_{i=1}^{I_{tt'}} \sum_{j=1}^{J-1} w_{i\ell s^*}(y_{itj} - \hat{\mu}_{itj})(y_{it'j} - \hat{\mu}_{it'j})}{\left\{\sum_{i=1}^{I_{tt'}} \sum_{j=1}^{J-1} w_{i\ell s^*}(y_{itj} - \hat{\mu}_{itj})^2\right\}^{\frac{1}{2}} \left\{\sum_{i=1}^{I_{tt'}} \sum_{j=1}^{J-1} w_{i\ell s^*}(y_{it'j} - \hat{\mu}_{it'j})^2\right\}^{\frac{1}{2}}}, \tag{4.4}$$

are consistent as $I_{tt'}$ are quite large in practice. For example, for the SLID data $I_{tt'}$ will be at least 16,890 with $T_0 = 2$, to compute α^*_{12}. Once the α_{tt}'s are computed, they are used in (4.2) and the equations are solved for β as usual. A consistent estimate of the covariance matrix of β^{**} (say) [solution of (4.2)] is then given by

$$\hat{\text{cov}}(\beta^{**}) = \left[\sum_{i \epsilon s^*}^{I} w_{i\ell s^*} W_{it_i}^T \Sigma_{it_i}^{-1}(\alpha^*) W_{it_i}\right]^{-1}. \tag{4.5}$$

4.2 Generalized estimating equations when data are missing at random (MAR)

In this section we deal with the regression analysis when data are missing at random. Note that as opposed to the MCAR case, in the MAR case the probability of missingness depends on the outcome. Consequently, $E\{r_{it}(y_{itj} - \mu_{itj})\} \neq 0$ and the root of the estimating equations (4.1) or (4.2) will no longer be unbiased estimate of β. Robins, Rotnitzky, and Zhao (1995) proposed a weighting method for rendering GEE analyses correct under missing-at-random (MAR) missing mechanisms. More specifically, these authors proposed to use the weighted estimating equations

$$\sum_{i \epsilon s^*} w_{i\ell s^*} W_i^{c^T} \Sigma_i^{c^{-1}}(\alpha) \Delta_i (Y_i^c - \mu_i^c) = 0, \tag{4.6}$$

where the tth diagonal element of Δ_i is $r_{it} / \Pr\left[\left\{\Pi_{u=1}^{t} r_{iu}\right\} = 1 | D_{i,t-1}\right]$, with $D_{i,t-1}$ is the history of the ith person up to time $t-1$; that is, $D_{i,t-1} = (x_i, Y_{i1}^{*T}, \ldots, Y_{i,t-1}^{*T})$, where Y_{i1}^*, for example, is the column vector of $(J-1)$-dimensional response vector collected at time $t = 1$, and $t = 1, \ldots, t_i$. Robins et al (1995) showed that if Δ_i is estimated consistently, then the root β_W^* is consistent and asymptotically normal under MAR.

More recently, Paik (1997) has suggested a mean imputation technique to remove the bias of the regression estimate obtained from (4.1). More specifically, Paik has suggested to replace missing Y_{it}^*

with $\tilde{y}_{it}$, an estimate of $E\{Y_{it}^*|D_{it_i}, r_{it}=0\}$ for $t = t_i+1, \ldots, T_0$. The resulting imputed estimating equations are

$$\sum_{i\epsilon s^*} w_{i\ell s^*} W_i^{c^T} \Sigma_i^{c^{-1}}(\alpha)(\tilde{Y}_i^c - \mu_u^c) = 0, \tag{4.7}$$

where $\tilde{Y}_i^c = (Y_{i1}^{*^T}, \ldots, Y_{it_i}^{*^T}, \tilde{y}_{i,t_i+1}^{*^T}, \ldots, \tilde{y}_{iT_0}^{*^T})^T$. If $(\tilde{y}_{i,t_i+1}^{*^T}, \ldots, \tilde{y}_{iT_0}^{*^T})^T$ is consistent for missing $(Y_{i,t_i+1}^{*^T}, \ldots, Y_{iT_0}^{*^T})^T$, then (4.7) is an unbiased estimating equation. This is because

$$u_{it} = r_{it}(Y_{it}^* - \mu_{it}) + (1-r_{it})\{E(Y_{it}^*|D_{it_i}, r_{it}=0) - \mu_{it}\}$$

has zero expectation vector, where the expectation is taken over the joint distribution of $(Y_{it}^{*^T}, r_{it}, Y_{i,t-1}^{*^T}, r_{i,t-1}, \ldots, Y_{i1}^{*^T}, r_{i1})$. The problem then reduces to estimating $E(Y_{it}^*|D_{it_i}, r_{it}=0)$, which

depends on the missingness probabilities. Paik (1997) has considered two MAR mechanisms, where it is assumed that the first time response cannot be missing which seems reasonable in a longitudinal study. For $t = 2, \ldots, T_0$, Paik's two MAR mechanisms are:

M1. $\Pr(r_{it}=1|Y_i^c, x_i, r_{i,t-1}=1) = \Pr(r_{it}=1|Y_{i1}^*, x_i, r_{i,t-1}=1)$

M2. $\Pr(r_{it}=1|Y_i^c, x_i, r_{i,t-1}=1) = \Pr(r_{it}=1|Y_{i1}^*, \ldots, Y_{i,t-1}^*, x_i, r_{i,t-1}=1)$.

Here, according to the first mechanism , the non-response probability depends only on the first observed data. Similarly, the second mechanism refers to the case where the response or non-response probability depends on all previous observed data.

Following Paik (1997), we now summarize the estimation of the missing values suitable for SLID data under any of the above two MAR mechanisms. In the context of SLID data, we assume that there is no missing individual in 1993 as it is the first survey following the selection of longitudinal units. To estimating a missing response for 1994, we use

$$\tilde{y}_{i2} = \tilde{E}(Y_{i2}^*|D_{i1}, r_{i1}=1, r_{i2}=0), \tag{4.8}$$

which, by application of Baye's theorem, is the same as

$$\tilde{y}_{i2} = \tilde{E}(Y_{i2}^*|D_{i1}, r_{i1}=1, r_{i2}=1) \tag{4.9}$$

under mechanisms M_1 and M_2. This implies that the expectation in (4.9) can be consistently estimated by the sample mean of observed $Y^*_{i't}$ vectors having the same history as D_{i1}; that is,

$$\tilde{y}_{i2} = \frac{\sum_{i'=1}^{I} Y^*_{i'2} r_{i'2} I(D_{i'1} = D_{i1})}{\sum_{i'=1}^{I} r_{i'2} I(D_{i'1} = D_{i1})}. \tag{4.10}$$

Next, we proceed to estimate the possible missing values for 1995, which should be useful in the near future once 1995 data is available for analysis. There are two situations to consider here. First, suppose that responses for 1993 and 1994 are available but 1995 information has to be imputted for a given individual. Under the above two mechanisms, this response may also be estimated by a formula similar to (4.9). More specifically,

$$\tilde{y}_{i3} = \frac{\sum_{i'=1}^{I} Y^*_{i'3} r_{i'3} I(D_{i'2} = D_{i2})}{\sum_{i'=1}^{I} r_{i'3} I(D_{i'2} = D_{i2})}. \tag{4.11}$$

In the second case, the response for 1993 is considered to be known and the remaining two responses have to be estimated. In this case, the missing response for 1994 is computed by (4.9). A consistent estimate for the missing response of 1995 is given by [cf. Paik (1997, eqn (5), p. 1321)]

$$\tilde{\tilde{y}}_{i3} = \frac{\sum_{i'=1}^{I} Y^*_{i'3} r_{i'3} I(D_{i'1} = D_{i1}) + \sum_{i'=1}^{I} \tilde{y}_{i3} r_{i'2} (1 - r_{i'3}) I(D_{i'1} = D_{i1})}{\sum_{i'=1}^{I} r_{i'2} I(D_{i'1} = D_{i1})}. \tag{4.12}$$

These missing value estimates are then used in (4.7) to obtain a consistent estimate of β. In this case, however, α in (4.7) is estimated by using (4.3) based on all available and imputted responses. Call this estimate as $\tilde{\alpha}$, and β estimate as $\tilde{\beta}$.

Next, a consistent estimate of the covariance matrix of $\tilde{\beta}$ is given by

$$\hat{\text{cov}}(\tilde{\beta}) \quad = \quad \left[\sum_{i \epsilon s^*} w_{i\ell s^*} W_i^{c^T} \Sigma_i^{c^{-1}}(\tilde{\alpha}) W_i^c \right]^{-1}, \tag{4.13}$$

which may be used to test the significance of the covariates in explaining the responses.

5 Concluding Remarks

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. Here, one is interested to examine the effects of the associated covariates on one or more suitable response variables, after taking the longitudinal correlations into account. In the present report, I have developed a multivariate regression approach to deal with such longitudinal data. More specifically, a multivariate approach was developled to analyse ordinal polytomous longitudinal survey data, such as SLID data collected by Statistics Canada. In this report, I have applied this methodology to the SLID data to examine the effects of covariates such as age, sex, marital status, race, education level and province of residence on the 'jobless spell' response variable. Here, jobless spells were partitioned into several ordinal categories and their probabilities were modelled by introducing a cumulative logistic approach which accommodates the order restrictions in a natural way. The longitudinal correlations between responses for 1993 and 1994 (for the SLID data) were modelled following Lipsitz and Fitzmaurice (1996) and estimated following Sutradhar and Das (1999). Details about the regression effects are reported in Section 3 of this report.

Note, however, that the regression analysis reported in Section 3 requires the assumption that there is no missing information over the duration of the survey. Consequently, a suitable sub-set of the SLID data was formed consisting of 16890 individuals with complete responses both for 1993 and 1994, and the numerical results reported in Section 3 were computed based on the complete information available from these 16890 individuals. Now, the dealing with a more larger data set (even with the complete/full data set) requires suitable modifications to the regression approach in order to obtain consistent regression estimates in the presence of missing values. As this was felt as an immediate important issue, in Section 4, I have developed a suitable modified estimating equations approach for MCAR and MAR types of missing data. This I have done following Paik (1997). The present methodology, however, unlike Paik (1997), provides a technique to obtain consistent estimates for the longitudinal correlations in the presence of missing values, which are then used in the estimating equations to obtain the regression estimates. It is shown how the regression estimates may be computed to deal with SLID data up to 1995, while the standard regression estimates reported in Section 3 were computed based on the complete available data for

1993 and 1994.

With regard to the appropriateness of the polytomous logistic model used in Section 3, it should be noted that some authors (cf. Williamson et al (1995), Molenburghs and Lesaffre (1994)) have modelled the ordinal multinomial probability model by introducing suitable cut points. These cut points however are estimated without challenging the order restricted nature of the cut points for ordinal categories which may yield inconsistent estimates. The propsed model is free from this limitation as the cumulative probability model is constructed based on cumulative logistic probabilities which takes care of the cut points problem in a natural way. Further, to deal with the longitudinal correlations, we have not used any 'working' correlation approach, rather we modelled the correlation structure in a robust way following Sutradhar and Das (1999). Therefore, the longitudinal approach used in the report to deal with polytomous ordinal data is quite robust as compared to the techniques available to deal with such logitudinal ordinal data.

Further note that the longitudinal ordinal model constructed in part I has been extended in this report to accomodate possible longitudinal missing responses. As mentioned earlier, this further modelling however requires some knowledge about the missingness nature of the data. Consequently, missingness misspecification may affect the regression estimates. A study to examine the robustness of missing values estmation technique used in the report would be worth considering, which is, however, beyond the scope of the present report.

With regard to the robustness of the models, further problem is mounted when longitudinal data are collected based on suitable complex survey design. In both part I and part II of the report, we have applied the longitudinal survey weights to construct a suitable estimator for a vector of finite population totals. These totals are functions of fixed regression parameters and these functions are solved for regression estimates by equating them to zero, that is, solving associated estimating equations. Now checking the appropriateness of the survey desin or more generally to check the robustness of the functions of finite population totals and their survey weights based estimators to solve for regression effects appear to be a difficult issue, which may be pursued in a separate study.

Acknowledgement

I thank Dr. Milorad Kovacevic for providing the SLID data discussed in the report. I also thank him for some useful suggestions and for his help in exploratory data analysis.

References

Kenward, M. G., Lesaffre, E., and Molenburghs, G. (1994). An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random, *Biometrics* 50, 945-953.

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

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

Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random, *J. Amer. Statist. Assoc.* 92, 1320-1329.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *J. Amer. Statist. Assoc.* 90, 106-121.

Sutradhar, B. C. (1998). Analysis of multivariate ordinal data for longitudinal surveys: Part I, *A report for Statistics Canada.*

Sutradhar, B. C., and Das, K. (1999). On the efficiency of regression estimators in generalized linear models for longitudinal data, *Biometrika*, Vol. 86, No. 2, to appear.

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.