Survey Methodology, December 1998
Vol. 24, No. 2, pp. 131-138
Statistics Canada

Longitudinal Analysis of Swiss Labour Force Survey Data by Multivariate Logistic Regression

PAUL-ANDRÉ SALAMIN[1]

ABSTRACT

In longitudinal surveys, simple estimates of change, such as differences of percentages may not always be efficient enough to detect changes of practical relevance, especially in sub-populations. The use of models, which can represent the dependence structure of the longitudinal survey, can help to solve this problem. One of the main characteristics observed by the Swiss Labour Force Survey (SLFS) is the employment status. As the survey is designed as a rotating panel, the data from the SLFS are multivariate categorical data, where a large proportion of the response profiles are missing by design. The multivariate logistic model, introduced by Glonek and McCullagh (1995) as a generalisation of logistic regression, is attractive in this context, since it allows for dependent repeated observations and incomplete response profiles. We show that, using multivariate logistic regression, we can represent the complex dependence structure of the SLFS by a small number of parameters, and obtain more efficient estimates of change.

KEY WORDS: Longitudinal binary data; Multivariate logistic model; Labour force survey.

1. INTRODUCTION

One of the main objectives of the Swiss Labour Force Survey (SLFS), is to produce estimates of change for the percentages of the population in different employment statuses. Typically, simple estimates of change, such as the difference of the percentages of employed individuals between two years, are calculated for the whole population, and for a large number of sub-populations. In general, this is unsatisfactory, as the estimates for the sub-populations may not always be efficient enough to detect changes of practical relevance. The work presented here was motivated by the question, whether the use of models, which can represent the dependence structure of the survey, could help to solve this problem.

As the SLFS is designed as a rotating panel, we are dealing with longitudinal categorical data, for which a fairly large proportion of the response profiles, are incomplete by design. The focus of interest is on modelling marginal probabilities, namely, the probabilities to be in a given employment status, as a function of time and other covariates that define sub-populations. If the repeated observations of the employment status were independent, a natural approach would be to use logistic regression. The multivariate logistic model, introduced by Glonek and McCullagh (1995) as a generalisation of logistic regression, is attractive in this context, since it allows for dependent repeated observations and incomplete response profiles.

The aim of this paper is to show that, the ability of multivariate logistic regression to model the complex dependence structure of the SLFS data, leads to more efficient estimators of change. Although we illustrate the method using the SLFS data only, it is clearly of wider applicability.

There are a number of important issues that are not dealt with in this paper. As the SLFS data come from a complex survey, it can be argued that any analysis should take the sampling weights into account (Pfeffermann 1993). Here we use the unweighted data only. However, it can be shown, using the pseudo-likelihood approach of Binder (1983), that multivariate logistic regression can be extended to that situation (Salamin 1998). Non-response is always of great concern in sample surveys. Here, we consider only the incomplete response profiles that arise through the rotation of the panel, in which case, the hypothesis of missing completely at random, is reasonable. Note however, that multivariate logistic regression, is flexible enough to incorporate extra parameters for the incomplete profiles, arising from panel, attrition. Thus, the individuals which dropped out of the panel, could also have been included into the analysis. Finally, it is well known that classification errors may introduce large biases in the observed response profile probabilities, see *e.g.*, Pfeffermann, Skinner and Keith (1998). It would certainly be desirable to investigate how these biases affect the parameter estimates of multivariate logistic regression, which have interpretations in terms of marginal moments.

Log-linear models and marginal models are closely related to multivariate logistic regression, and are further discussed in Section 3. Here we discuss briefly transition models, random effects models, and survival analysis, in the context of the SLFS. Under a transition model, see *e.g.*, Diggle, Liang and Zeger (1994, Ch. 10) or Zeger and Liang (1992), the repeated observations of the employment status are correlated, because past employment statuses influence the present employment status. The focus of interest, are the transition probabilities between the different employment statuses, *e.g.*, the probability of being

[1] Paul-André Salamin, Statistical Methods Unit, Swiss Federal Statistical Office, Espace de l'Europe 10, CH-2010 Neuchâtel, Switzerland.

employed, conditional on being unemployed in the past. In the regression setting, the past responses are treated as additional explanatory variables. An important issue, is the determination of the number of past responses to include as predictors. If the model for the transition probabilities is correctly specified, we can treat the repeated transitions for an individual as independent events, and use standard statistical methods, such as logistic regression. Under a random effect model, see *e.g.*, Diggle *et al.* (1994, Ch. 9), the probability of being in a given employment status, is a function of explanatory variables, where the regression coefficients vary from one individual to the next. This variability of the regression coefficients, reflects the natural heterogeneity of the individuals, due to unmeasured factors. Given the regression coefficients, the repeated observations of the employment status, are assumed to be independent. The correlation among the repeated observations, arises solely because we are unable to observe the true regression coefficients. This approach is most useful, when inference about individuals rather than population averages, is the focus of interest. In survival analysis, also called event history analysis in the econometric literature (Lancaster 1990), the focus is on modelling the transitions between employment statuses over time, as a function of explanatory variables. Here, the exact time at which a transition takes place, is important. In the SLFS, the employment status is observed once a year. The changes in employment status, that took place during the year preceding the interview, can be reconstructed. However, since this reconstruction is based on the self-assessment of the subjects, there may be some imprecision as regards prior status, and time of change of status. An analysis of the SLFS data based on this approach can be found in Gerfin (1996).

The article is organized as follows. We begin in Section 2 by describing the data, a subset of about 5000 individuals from the SLFS, which are used in the examples of Sections 4 and 5. In Section 3, we discuss multivariate logistic regression, and contrast it with the log-linear and marginal models. In Section 4, we illustrate the ability of multivariate logistic regression, to represent the complex dependence structure of the SLFS data, by a small number of parameters. In Section 5, we compare multivariate logistic regression with a simple estimator of change. It is shown that, using multivariate logistic regression, results in a gain in efficiency. Finally, we present in Section 6 our conclusions, and give directions for further work.

2. SWISS LABOUR FORCE SURVEY DATA

A detailed description of the sampling design and weighting procedure of the SLFS, can be found in Hulliger, Ries, Comment and Bender (1997). Here, we just recall some of the relevant aspects of this survey. The SLFS collects information on the employment of resident persons of age 15 or more in Switzerland. Starting in the second quarter of 1991, a sample of about 16,000 persons are interviewed each year. The survey is designed as a rotating panel, with a time-in-sample of 5 years. During the start-up phase, *i.e.*, from 1992 to 1996, approximately one fifth of the original sample was rotated out each year, and replaced by a renewal sample. The units in the renewal samples then stayed in the panel for a full period of 5 years.

In the examples of Sections 4 and 5, we use the observations of the employment status, for the years 1992 to 1995, obtained from the individuals in the sample, of the canton of Vaud. The structure of the data, as well as the longitudinal and cross-sectional sample sizes, are shown in Table 1. Due to the sampling design, some of the response profiles are incomplete. For example, for the individuals that were selected in 1991 and rotated out of the sample in 1994, the period of observation, denoted (1)234, goes from 1991 to 1994. We use the notation (1)234, to emphasise the fact, that we do not use the observations taken in 1991.

Table 1
Structure of the Data, Longitudinal and Cross-sectional Sample Sizes Canton of Vaud, 1992-1995

First year in sample	Observation times for various parts of the sample				Period of observation	
91	92				(1)2	622
	92	93			(1)23	412
	92	93	94		(1)234	527
	92	93	94	95	(1)2345	481
92	92	93	94	95	2345	612
93		93	94	95	345	722
94			94	95	45	728
95				95	5	877
	2,654	2,754	3,070	3,420		4,981

Employment status is a nominal variable with three categories, defined as "employed", "unemployed" and "out of the labour force". In the examples of Section 4 and 5, we work with a binary variable, taking the value 1 if an individual is employed, and 2 if an individual is unemployed or out of the labour force. This is done solely to simplify the presentation of the multivariate logistic models. As the method can handle an arbitrary number of categories, it would be preferable, not to collapse the statuses in a real analysis. Caution must be exercised, if it is nevertheless necessary to combine some of the statuses, as heterogeneity of the statuses may introduce bias.

3. MULTIVARIATE LOGISTIC MODELS

The multivariate logistic model, introduced by Glonek and McCullagh (1995), can handle multivariate responses of either nominal or ordinal types, and either discrete or continuous explanatory variables. Here, we consider only multivariate binary responses and discrete predictors. The multivariate logistic model, is an example of a generalized linear model, see McCullagh and Nelder (1989). Its link function, also called the multivariate logistic transformation, expresses the joint distribution of the response profiles, in terms of marginal moments of increasing order, the first two being marginal logits, and marginal log odds ratios. The link function has the property, termed reproducibility, that a multivariate logistic model, applies to any subset of the response vector. This property ensures that, the interpretations of the parameters are the same, regardless of the number of response variables, and whether or not higher order parameters are included. This makes multivariate logistic regression, especially attractive for the analysis of longitudinal data, where the repeated observations of an outcome arise on an equal footing, and where the number of repeated observations may vary from one individual to the next. Reproducibility is also the key to the ability of the model, to accommodate observations with incomplete responses. Note however, that we need to assume, that the data are missing completely at random, if the same parameters are to be used to model the complete and incomplete response profiles. The parameter estimates are found by maximum likelihood. A key step, is the inversion of the multivariate logistic transformation. For more than three responses, this may not always be possible, as there are then constraints among the parameters (Glonek and McCullagh 1995, Liang, Zeger and Qaqish 1992). Also, the presence of empty cells, may limit the order of the parameters that can be fitted.

The log-linear model is widely used to model multivariate binary data. In the saturated log-linear model, see *e.g.*, Liang *et al.* (1992), the canonical parameter associated with a subset of the variables, has an interpretation in terms of conditional probabilities given the rest of the variables, *e.g.*, the first and second order parameters are logits and log odds ratios, conditional on all the other responses. It follows that, the log-linear model is not reproducible, which makes it less preferable than multivariate logistic regression, for the analysis of longitudinal data. It is nevertheless possible, to build log-linear models that, as in the multivariate logistic model, have marginal logits as parameters. This leads to the marginal models (Diggle *et al.* 1994, Ch. 8). In these models, the dependence of the marginal probabilities on explanatory variables, is modelled separately from within-unit correlation. Under this approach, the parameters are not estimated by maximum likelihood. Rather, only the structure of the correlation, between the repeated observations of an outcome is specified, and the parameters are estimated by solving generalized estimating equations (GEE), a multivariate analogue of quasi-likelihood (McCullagh and Nelder 1989). A number of specifications of the correlation structure have been proposed, for example Liang *et al.* (1992) use the marginal log odds ratios, as in Glonek and McCullagh (1995). We have made some comparisons between multivariate logistic regression and PROC GENMOD of SAS (release 6.12). This procedure has the ability to fit correlated response models by the GEE method. We found very similar estimates of the marginal logits. The GEE method appeared to be slightly less efficient than multivariate logistic regression. A limitation of the GEE method is that, it cannot yield estimates of the response profile probabilities, but only of the marginal probabilities. By contrast, the multivariate logistic model does not have this limitation, since its parameters are estimated by maximum likelihood.

Following Glonek and McCullagh (1995), we discuss in Section 3.1 the multivariate logistic transformation, and we give, in Section 3.2, the algorithm for maximum likelihood.

3.1 Multivariate Logistic Transformation

Let $Y_1, Y_2, ..., Y_d$ be d repeated observations, taken at times $t_1 < t_2 < ... < t_d$, of the same binary variable, and let

$$\pi_{i_1 i_2 ... i_d} = P(Y_1 = i_1, Y_2 = i_2, ..., Y_d = i_d),$$

where $i_1, i_2, ..., i_d$ are all either 1 or 2, be the joint probabilities of the random variables $Y_1, Y_2, ..., Y_d$. In the multivariate logistic model, the joint probabilities of $Y_1, Y_2, ..., Y_d$ are parameterized in terms of marginal logits, marginal log odds ratios, and contrasts of marginal log odds ratios. This parameterization can be written as $\eta = C^T \log(L\pi)$, where π is the vector of dimension $q = 2^d$

$$\pi = (\pi_{11...11}, \pi_{11...12}, ..., \pi_{22...21}, \pi_{22...22})^T,$$

and where, the matrices L and C are tensor products of suitably chosen marginal indicator and contrast matrices respectively. The matrices L and C, which depend on the length d of the observation period, are defined recursively, beginning with $L_0 = C_0 = 1$, as

$$L_d = \begin{bmatrix} L_{d-1} \otimes 1_2^T \\ L_{d-1} \otimes \tilde{L} \end{bmatrix}$$

and

$$C_d = \begin{bmatrix} C_{d-1} & 0 \\ 0 & C_{d-1} \otimes \tilde{C} \end{bmatrix},$$

where $1_2^T = (1, 1)$, $\tilde{L}$ is the two by two identity matrix, and $\tilde{C} = (1, -1)^T$ (Glonek and McCullagh 1994).

To illustrate matters, we consider periods of observation of length $d = 1, 2, 3, 4$. For $d = 1$, $\pi = (\pi_1, \pi_2)^T$ and $\eta = (\eta_0, \eta_1)^T = (\log \pi_+, \text{logit } Y_1)^T$, where the plus subscript indicates summation, and logit Y_1 is defined as

$$\text{logit } Y_1 = \log \frac{P(Y_1 = 1)}{P(Y_1 = 2)} = \log \frac{\pi_1}{1 - \pi_1} = \log \frac{\pi_1}{\pi_2}.$$

In that case the multivariate logistic transformation is equivalent to the usual logistic transformation. Note that, although the parameter $\eta_0 = \log \pi_+ = 0$ is strictly superfluous, it is convenient to retain it, as a means of ensuring that the mapping $\pi \rightarrow \eta$ is of full rank, and also expressing the requirement that $\pi_+ = 1$.

For $d = 2, \pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})^T$ and

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12})^T =$$

$$(\log \pi_{++}, \text{logit } Y_1, \text{logit } Y_2, \log OR(Y_1, Y_2))^T$$

where

$$OR(Y_1, Y_2) =$$

$$\frac{P(Y_1 = 1, Y_2 = 1)\, P(Y_1 = 2, Y_2 = 2)}{P(Y_1 = 1, Y_2 = 2)\, P(Y_1 = 2, Y_2 = 1)} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

is the odds ratio, a quantity measuring the association between the variables Y_1 and Y_2. The parameters η_1 and η_2 are the marginal logits at times t_1 and t_2, for example

$$\eta_1 = \text{logit } Y_1 = \log \frac{\pi_{1+}}{(1 - \pi_{1+})}.$$

For $d = 3, \pi = (\pi_{111}, \pi_{112}, \ldots, \pi_{221}, \pi_{222})^T$ and

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12}, \eta_3, \eta_{13}, \eta_{23}, \eta_{123})^T.$$

The parameters η_1, η_2 and η_3 are the marginal logits at times t_1, t_2 and t_3. The parameters η_{12}, η_{13} and η_{23} are the log odds ratios of the corresponding two-dimensional marginal tables, for example

$$\eta_{23} = \log OR(Y_2, Y_3) = \log \frac{\pi_{+11}\, \pi_{+22}}{\pi_{+12}\, \pi_{+21}}.$$

The parameter η_{123} is a contrast of log odds ratios given by

$$\eta_{123} = \log OR(Y_1, Y_2 \mid Y_3 = 1) - \log OR(Y_1, Y_2 \mid Y_3 = 2)$$
$$= \log \frac{\pi_{111}\pi_{221}}{\pi_{121}\pi_{211}} - \log \frac{\pi_{112}\pi_{222}}{\pi_{122}\pi_{212}}.$$

For $d = 4, \pi = (\pi_{1111}, \pi_{1112}, \ldots, \pi_{2221}, \pi_{2222})^T$ and

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12}, \eta_3, \eta_{13}, \eta_{23}, \eta_{123},$$
$$\eta_4, \eta_{14}, \eta_{24}, \eta_{124}, \eta_{34}, \eta_{134}, \eta_{234}, \eta_{1234})^T.$$

The parameters η_i, η_{ij} and η_{ijk}, where $1 \le i < j < k \le 4$, are defined as above, using the appropriate marginal tables. The parameter η_{1234} is a contrast of log odds ratios given by

$$\eta_{1234} = \log OR(Y_1, Y_2 \mid Y_3 = 1, Y_4 = 1)$$
$$- \log OR(Y_1, Y_2 \mid Y_3 = 1, Y_4 = 2)$$
$$- \log OR(Y_1, Y_2 \mid Y_3 = 2, Y_4 = 1)$$
$$+ \log OR(Y_1, Y_2 \mid Y_3 = 2, Y_4 = 2).$$

A key step in maximum likelihood estimation is the computation of the inverse of the multivariate logistic transformation. To ensure that $\pi > 0$, we work with $\pi = \exp v$, *i.e.*, we seek to solve for v in the equation $\eta = C^T \log(L \exp v)$. In general, no explicit solution is available, so an iterative method must be used. In particular, the Newton-Raphson iterations can be applied as described below. For clarity, we define the two functions $\varphi(\pi) = C^T \log(L\pi)$ and $\psi(v) = \varphi(\exp v)$.

(i) Begin with an initial approximation v_0.

(ii) Then take $v_{k+1} = v_k - [D\psi(v_k)]^{-1}(\varphi(\exp v_k) - \eta)$, where $D\psi(v)$ is the Jacobian matrix of the function $\psi(v)$, and iterate until convergence.

The Jacobian matrices of the function $\varphi(\pi)$ and $\psi(v)$ are given respectively by $D\varphi(\pi) = C^T (\text{diag} L\pi)^{-1} L$ and $D\psi(v) = D\varphi(\exp v) \cdot \text{diag}(\exp v)$.

3.2 Maximum Likelihood Estimation

For a binary response variable observed at d time points, there are $q = 2^d$ possible response profiles $i = (i_1, \ldots, i_d)$, where $i_1, i_2, \ldots, i_d$ are all either 1 or 2. For each profile $i = (i_1, \ldots, i_d)$, we define the indicator variable $Y_{i_1 \ldots i_d}$, which is equal to 1 if the profile i has been observed, and 0 otherwise. We then have

$$P(Y_{i_1 \ldots i_d} = 1) = P(Y_1 = i_1, \ldots, Y_d = i_d) = \pi_{i_1 \ldots i_d}.$$

Defining the q-dimensional vectors

$$Y = (Y_{11\ldots11}, Y_{11\ldots12}, \ldots, Y_{22\ldots21}, Y_{22\ldots22})^T$$

and

$$\pi = (\pi_{11\ldots11}, \pi_{11\ldots12}, \ldots, \pi_{22\ldots21}, \pi_{22\ldots22})^T,$$

we may then write $Y \sim M(1, \pi)$, *i.e.*, Y is a multinomial vector with $q = 2^d$ categories, whose probabilities are given by the vector π.

The multivariate logistic regression models, are then defined to be those of the form $\eta = X\beta$ where X is a $q \times p$ matrix of explanatory variables, β is a p-dimensional vector of unknown parameters, and $\eta = C^T \log(L\pi) = \varphi(\pi)$.

If we let y be one observation of the random vector Y, then we may write the kernel of the log likelihood function as $l(\beta; y) = y^T \log \pi(\beta)$ where, using the inverse of the multivariate logistic transformation, we can express the joint probabilities π as a function of the unknown parameter β, as $\pi(\beta) = \varphi^{-1}(X\beta)$. The score vector is given by

$$s(\beta) = s(\beta, y, X) = D\pi(\beta)^T (\operatorname{diag}\pi(\beta))^{-1} y,$$

where $D\pi(\beta)$, the Jacobian matrix of the function $\pi(\beta)$, relating the parameter β to the vector of probabilities π, is given by $D\pi(\beta) = [D\varphi(\varphi^{-1}(X\beta))]^{-1}X$, and where $D\varphi(\pi) = C^T(\operatorname{diag} L\pi)^{-1}L$, is the Jacobian matrix of the link function. The information matrix is defined as $\mathfrak{I}(\beta) = Es(\beta)s(\beta)^T$. Now it follows from the assumption on the distribution of Y, that $E(YY^T) = \operatorname{diag}\pi$, from which we may deduce that

$$\mathfrak{I}(\beta) = \mathfrak{I}(\beta, X) = D\pi(\beta)^T (\operatorname{diag}\pi(\beta))^{-1} D\pi(\beta).$$

If we have n independent observations $y_k \sim M(1, \pi_k)$, $k = 1, ..., n$, where $\eta_k = C^T \log(L\pi_k) = X_k\beta$, then the score vector and the information matrix are given by $s(\beta) = \sum_{k=1}^n s(\beta, y_k, X_k)$ and $\mathfrak{I}(\beta) = \sum_{k=1}^n \mathfrak{I}(\beta, X_k)$.

The maximum likelihood estimator of β is the solution of $s(\beta) = 0$, that can be found by using the Fisher scoring algorithm which, starting from some initial value β_0, iterates the sequence $\beta_{m+1} + \beta_m + \mathfrak{I}_m^{-1}(\beta_m)s(\beta_m)$ until convergence.

Incomplete response profiles can readily be incorporated into the analysis. In particular, if some subset of the response variables $Y_1, Y_2, ..., Y_c$ is recorded for a particular unit, then the probability distribution on that c-dimensional marginal table is multinomial, and, as a consequence of the reproducibility of the multivariate logistic transformation, a multivariate logistic regression model applies to the table of probabilities. Furthermore, the design matrix relating the marginal probabilities to β, is constructed by selecting the appropriate rows of the full design matrix, that would be used if complete data were available for that unit.

4. MODELS FOR LONGITUDINAL DEPENDENCE

In this section we illustrate, using the SLFS data of Section 2, how multivariate logistic regression can be applied to describe the dependence between the repeated observations of the employment status. We do not intend to carry out an exhaustive search for a best model, but rather to demonstrate the ability of the method, to represent a complex dependence structure by a small number of parameters.

We consider 6 models of decreasing complexity, see Table 2. For all 6 models, we have one parameter for each of the marginal logits corresponding to a given observation time. Symbolically, this is denoted by $\eta_i = \beta_i$. Since the observation times are the 2nd quarter of the years 1992 to 1995, we take $i = 2, 3, 4, 5$. Thus β_3, say, corresponds to the logit of the probability of being employed in 1993. Similarly, the indices for the higher order parameters run from 2 to 5. For model 1 we take a saturated model for the longitudinal dependence, *i.e.*, we have one parameter for each of the interactions of order 2, 3 or 4 within each period of observation. For the models 2 to 5, we assume that the interactions of order 3 and 4, are all equal to zero. The longitudinal dependence is then described in terms of log odds ratios only. For model 2, we take a saturated model for the log odds ratios. In model 3 we drop the covariate period of observation, *i.e.*, we suppose that the log odds ratios are the same for all the periods of observation. In model 4, we use stationary log odds ratios, *i.e.*, log odds ratios which depend only on the difference between times of observation. Note that the parameter γ_1 in model 4, corresponds to the constraint $\beta_{23} = \beta_{34} = \beta_{45}$ on the parameters of model 3, and similarly for γ_2 and γ_3. In model 5, a linear model for the stationary log odds ratios is assumed. In model 6, finally, we assume that the observations taken at different times, are independent. Note that, in that case, multivariate logistic regression is equivalent to ordinary logistic regression.

Table 2
Six Models for Longitudinal Dependence

	Parameters		
Model	Marginal logits	Log odds ratios	3rd and 4th order parameters
1	$\eta_i = \beta_i$	$\eta_{ij} = \beta_{ij,\text{period}}$	$\eta_{ijk} = \beta_{ijk,\text{period}}$, $\eta_{ijkl} = \beta_{ijkl,\text{period}}$
2	$\eta_i = \beta_i$	$\eta_{ij} = \beta_{ij,\text{period}}$	$\eta_{ijk} = 0, \eta_{ijkl} = 0$
3	$\eta_i = \beta_i$	$\eta_{ij} = \beta_{ij}$	$\eta_{ijk} = 0, \eta_{ijkl} = 0$
4	$\eta_i = \beta_i$	$\eta_{ij} = \gamma_{\lvert i-j \rvert}$	$\eta_{ijk} = 0, \eta_{ijkl} = 0$
5	$\eta_i = \beta_i$	$\eta_{ij} = \delta + \gamma . \lvert i - j \rvert$	$\eta_{ijk} = 0, \eta_{ijkl} = 0$
6	$\eta_i = \beta_i$	$\eta_{ij} = 0$	$\eta_{ijk} = 0, \eta_{ijkl} = 0$

The parameter estimates for the models 2 to 6, are given in Table 3. The number of parameters and the values of the log likelihood function at the maximum likelihood estimates, can be found in Table 4 where, for comparison, we also included the log likelihood for the fully saturated model.

Overall, we notice that the assumed form of the longitudinal dependence, appears to have little effect on the estimates of the marginal logits. This is a desirable property, as the marginal logits would typically be the parameters of interest. The standard errors of the marginal logits, are almost the same for the models that take into account the longitudinal dependence, but are inflated by about 15% for ordinary logistic regression (model 6). It can also be shown that the estimates of the marginal logits are positively correlated under the models that assume a longitudinal dependence, and uncorrelated for ordinary logistic regression. For the example considered here, the correlation was found to lie between 0.4 and 0.8. Thus, modelling the longitudinal dependence, leads also to more efficient estimates of the difference of marginal logits.

It can be seen from the fit of model 1, that the interaction parameters of order 3 and 4, are not significantly different from 0. This suggests that the longitudinal dependence can be described by the log odds ratios only. This hypothesis is corroborated by the incremental deviance of model 2 with respect to model 1, which is found to be 7.9, on 12 degrees of freedom. Further, all the parameters of model 2 are significantly different from 0, and an examination of the standardised residuals for the fitted probabilities of the response profiles, does not reveal any anomaly. For applications in official statistics, model 2 would be the preferred model, since it is based on as few assumptions as possible, while still allowing a substantial reduction in the number of parameters, thus rendering less acute the danger of sparse tables when longer periods of observation and models with more covariates are considered.

The models 3, 4 and 5 show that, it would nevertheless be possible to greatly simplify the description of the longitudinal dependence, without losing too much information. In going from model 2 to model 5, we observe that the deviance from the fully saturated model, does not increase much, see Table 4. Further, an examination of the residuals shows that, the models 3, 4 and 5 fit the data almost as well as model 2. On the other hand, while model 2 requires 20 parameters to describe the longitudinal dependence, model 5 needs only 2 parameters. This must be contrasted with model 6, which assumes independence between observations taken at different times: the log likelihood is much smaller than for the fully saturated model, see Table 4, and the fit to the data is poor.

Table 3
Parameter Estimates and Standard Errors

Parameter	Period	Model 2	Model 3	Model 4	Model 5	Model 6
logit 92		0.6348 (0.0350)	0.6360 (0.0352)	0.6348 (0.0352)	0.6347 (0.0352)	0.6471 (0.0409)
logit 93		0.5555 (0.0335)	0.5570 (0.0338)	0.5597 (0.0335)	0.5601 (0.0335)	0.5509 (0.0396)
logit 94		0.5440 (0.0324)	0.5407 (0.0325)	0.5402 (0.0326)	0.5397 (0.0325)	0.5377 (0.0374)
logit 95		0.4699 (0.0317)	0.4711 (0.0320)	0.4710 (0.0320)	0.4712 (0.0320)	0.4705 (0.0351)
β_{23}	(1)23	4.2563 (0.3311)	4.2579 (0.1465)			
	(1)234	4.2003 (0.2894)				
	(1)2345	4.0859 (0.2954)				
	2345	4.4830 (0.2841)				
β_{34}	(1)234	4.0894 (0.2794)	4.1111 (0.1310)			
	(1)2345	3.9611 (0.2840)				
	2345	4.0989 (0.2600)				
	345	4.2490 (0.2468)				
β_{45}	(1)2345	5.3992 (0.3854)	4.5561 (0.1389)			
	2345	3.9779 (0.2544)				
	345	4.7288 (0.2735)				
	45	4.5069 (0.2600)				
β_{24}	(1)234	3.7168 (0.2641)	3.8371 (0.1442)			
	(1)2345	4.2560 (0.3059)				
	2345	3.5330 (0.2370)				
β_{35}	(1)2345	4.4000 (0.3098)	3.7913 (0.1334)			
	2345	3.6493 (0.2396)				
	345	3.6116 (0.2192)				
β_{25}	(1)2345	4.3984 (0.3173)	3.5774 (0.1530)			
	2345	3.2209 (0.2256)				
γ_1				4.3260 (0.0928)		
γ_2				3.8519 (0.1050)		
γ_3				3.5340 (0.1495)		
δ					4.7341 (0.1266)	
γ					-0.4191 (0.0653)	

Table 4
Number of Parameters and Value of the Log Likelihood Function at the Maximum Likelihood Estimates

Model	Number of parameters of order					Log likelihood
	1	2	3	4	Total	
Full Model	20	20	10	2	52	-5342.7
1	4	20	10	2	36	-5345.4
2	4	20	0	0	24	-5349.4
3	4	6	0	0	10	-5365.2
4	4	3	0	0	7	-5368.9
5	4	2	0	0	6	-5369.5
6	4	0	0	0	4	-7815.3

5. COMPARISON WITH SIMPLE ESTIMATE OF CHANGE

In this section we concentrate on the estimation of the difference of the probabilities of being employed between any two given years. We show that, estimates based on multivariate logistic regression, are more efficient than simple estimates defined as the difference of the proportions of employed individuals.

The model considered here, is the model 2 of Section 4, with sex as an additional explanatory variable. We have, for each sex, one parameter for each of the marginal logits corresponding to a given year. The longitudinal dependence is accounted for by a saturated model for the log odds ratios. The third and fourth order parameters are set to 0. This model has therefore 8 parameters for the marginal logits, and 40 parameters for the log odds ratios: 2 sexes x 20 odds ratios within periods of observation, see Table 3. By inverting the multivariate logistic transformation, estimates of the probability of being employed, and of their differences between any two given years, can also be computed.

A simple estimator of change is given by the difference of the proportions of employed individuals between any two given years. Its variance, which takes into account the overlap of the two samples, is given by

$$\frac{1}{n+r}\pi_{1+}(1-\pi_{1+}) + \frac{1}{n+c}\pi_{+1}(1-\pi_{+1})$$
$$-2\frac{n}{(n+r)(n+c)}(\pi_{11}-\pi_{1+}\pi_{+1}),$$

where n is the number of cases for which observations are available for both years, r and c are the number of cases for which observations are available for only one year, π_{11} is the probability of being employed during both years, and π_{1+} and π_{+1} are the marginal probabilities of being employed.

Tables 5 shows, for the SLFS data of Section 2, the estimates of the difference of the probability of being employed under both methods. Note that both methods yield similar estimates of change. The standard errors of the simple estimates, are on the average, 30% larger than for multivariate logistic regression. The corresponding mean relative efficiency of multivariate logistic regression, with respect to the simple estimates, is 1.7. By comparison, the mean relative efficiency of multivariate logistic regression with respect to ordinary logistic regression, is 3.2.

Table 5
Change in the Probability of Being Employed
Canton of Vaud, 1992-1995

	Comparison	Multivariate logistic regression	Simple estimate
Woman	92 *vs.* 93	0.0138 (0.0090)	0.0136 (0.0115)
	92 *vs.* 94	0.0184 (0.0102)	0.0168 (0.0134)
	92 *vs.* 95	0.0375 (0.0109)	0.0356 (0.0149)
	93 *vs.* 94	0.0047 (0.0087)	0.0031 (0.0107)
	93 *vs.* 95	0.0238 (0.0095)	0.0219 (0.0128)
	94 *vs.* 95	0.0191 (0.0076)	0.0188 (0.0100)
Men	92 *vs.* 93	0.0220 (0.0095)	0.0283 (0.0116)
	92 *vs.* 94	0.0245 (0.0102)	0.0334 (0.0133)
	92 *vs.* 95	0.0387 (0.0106)	0.0452 (0.0144)
	93 *vs.* 94	0.0024 (0.0092)	0.0052 (0.0111)
	93 *vs.* 95	0.0167 (0.0098)	0.0169 (0.0130)
	94 *vs.* 95	0.0143 (0.0080)	0.0117 (0.0102)

6. CONCLUSIONS

The analyses of the SLFS data presented here, have shown the usefulness of multivariate logistic regression. Modelling the longitudinal dependence is necessary, in order to obtain a satisfactory fit of the observed response profile probabilities. Ignoring the longitudinal dependence, we still obtain acceptable point estimates of the marginal logits, but the information on the detailed structure of the data is lost. Modelling the longitudinal dependence leads also to more efficient estimates of the marginal parameters and of change, when compared with ordinary logistic regression, and a simple estimator of change. Finally, the ability of multivariate logistic regression to represent a complex dependence structure, by a small number of parameters, has also been illustrated.

Using the results of Glonek and McCullagh (1995), it is possible to extend the examples presented here, to multivariate responses of either nominal or ordinal types, with either discrete or continuous explanatory variables. The method can also be extended, to take the sampling weights into account (Salamin 1998). For the SLFS, it was found that the sampling weights have little effect on the parameter estimates of the multivariate logistic model. The standard error of the parameter estimates, was inflated by about 15%. This moderate increase of the variability of the parameter estimates due to the sampling weights, is plausible.

Indeed, as in the SLFS, only one person per household is selected, a large cluster effect was not expected.

Apart from the sort of analyses presented here, multivariate logistic regression may also be used for modelling non-response probabilities in longitudinal studies. Such models may be useful when the sampling weights need to be adjusted for non-response. The ability of multivariate logistic regression to give a parsimonious model of the data, may also be of interest in small-area estimation. In particular, estimators for a given geographical region could be based on models for an appropriately chosen larger region.

Although we did not encounter serious problems in the examples presented here, further work may need to be done on the problem of sparse tables. A critical step, when there are a large number of empty cells, is the inversion of the multivariate logistic transformation. The approach of Lang (1996), where the inversion of the link function is avoided, by specifying the models through constraints, may be of interest in this context. Another area of investigation is the influence of the classification errors on the parameter estimates of the multivariate logistic model.

ACKNOWLEDGEMENTS

This paper benefited from discussions with colleagues at the Swiss Federal Statistical Office, among which Beat Hulliger and Philippe Eichenberger deserve special mention. The help of Ariane Bender, from the Section responsible for the Swiss Labour Force Survey, in preparing and understanding the data, was extremely valuable. The author also thanks the editor and two referees for their excellent suggestions that have led to a significant improvement of this paper.

REFERENCES

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

DIGGLE, P.J., LIANG, K.-Y., and ZEGER, S.L. (1994). *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

GERFIN, M. (1996). *Entwicklung von ökonometrischen Modellen zur Analyse der Dynamik auf dem schweizerischen Arbeitsmarkt.* SLFS-News, Swiss Federal Statistical Office, Berne.

GLONEK, G.F.V., and McCULLAGH, P. (1994). Multivariate Logistic Models. Technical Report 94-31, School of Information Science and Technology, Flinders University of South Australia, Adelaide.

GLONEK, G.F.V., and McCULLAGH, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society*, B, 57, 533-546.

HULLIGER, B., RIES, A., COMMENT, T., and BENDER, A. (1997). Weighting the Swiss Labour Force Survey. (Eds. C. Malaguerra, S. Morgenthaler and E. Ronchetti). In *Conference on Statistical Science Honoring the Bicentennial of Stefano Franscini's Birth, Monte Verità, Switerland*, Basel: Birkhäuser Verlag.

LANCASTER, T. (1990). *The Econometric Analysis of Transition Data.* Cambridge: Cambridge University Press.

LANG, J.B. (1996). Maximum likelihood methods for a generalized class of log-linear models. *Annals of Statistics*, 24, 726-752.

LIANG, K.-Y., ZEGER, S.L., and QAQISH, B. (1992). Multivariate Regression Analysis for Categorical Data. *Journal of the Royal Statistical Society*, B, 54, 3-40.

McCULLAGH, P., and NELDER, J.A. (1989). *Generalized Linear Models*, (2nd edn.). London: Chapman and Hall.

PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

PFEFFERMANN, D., SKINNER, C., and KEITH, H. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society*, A, 161, Part 1, 13-32.

SALAMIN, P.-A. (1998). Multivariate logistic regression for data from complex surveys. To appear *Proceedings: Symposium '98, Longitudinal Anlysis for Complex Surveys*, Statistics Canada, May 1998.

ZEGER, S.L., and LIANG, K.-Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, 11, 1825-1839.