Biometrika (1994), **81**, 4, *pp.* 767–75
Printed in Great Britain

Miscellanea

Logistic regression for autocorrelated data with application to repeated measures

By A. AZZALINI

Dipartimento di Scienze Statistiche, Università degli Studi di Padova, Via S. Francesco, 33, 35121 Padova, Italia

Summary

A stochastic model is proposed for the study of the influence of time-dependent covariates on the marginal distribution of the binary response in serially correlated binary data. Markov chains are expressed in terms of transitional rather than marginal probabilities. We show how to construct the model so that the covariates relate only to the mean value of the process, independently of the association parameter. After formulating the stochastic model for a simple sequence of data with possibly missing data, the same approach is applied to a repeated measures setting and illustrated with a real data example.

Some key words: Correlated binary data; Discrete time series; Logistic regression; Longitudinal data; Markov chain; Missing data; Odds ratio; Repeated measures; Serial dependence.

1. Introduction

Much current literature is concerned with the analysis of binary data collected at successive time points to examine the relationship between the probability of success and some time dependent covariates. In the simplest case, a sequence $(y_1, \ldots, y_T)$ is collected at equally spaced time points, with a k-dimensional covariate x_t associated with time t $(t = 1, \ldots, T)$. We are also interested in multiple sequences, i.e. repeated measures analysis. A frequent complication in repeated measures settings is missing data.

Markov chains are an obvious model. There are however many distinct ways of introducing them. As remarked by Ware, Lipsitz & Speizer (1988), a first broad distinction is between transitional and marginal models, depending on whether the covariates determine the conditional distribution given the past or the marginal distribution of any nominated observation. Within transitional models, the effect of the covariates may be on the transition probabilities of the Markov chain or on its mean value; Stiratelli, Laird & Ware (1984), Zeger & Qaqish (1988) and Cox & Snell (1989, pp. 96–102) are in this framework. However, it seems to us more plausible that, after the model has been fitted, one would want to use it for predicting the mean value of the next response on the basis of the covariates alone. Thus we prefer marginal models.

There has recently been widespread use of generalised estimating equations (Liang & Zeger, 1986) to overcome the difficulties related to 'the lack of a rich class of models such as the multivariate Gaussian'. The key feature of their method is that one does not attempt to model the joint distribution of the subject profile: only the marginal distribution at each time point is modelled as a function of the covariates, and the standard errors of the regression parameters are adjusted to allow for data autocorrelation. See also Lipsitz, Laird & Harrington (1992). A recent publication containing an up-to-date list of references is Carey, Zeger & Diggle (1993). While this approach has much merit, it does not produce a model for the stochastic mechanism which generates the data. The present paper explores the possibility of developing proper statistical models for some of the situations for which generalised estimating equations provide a solution. As a by-product, we obtain methods applicable to single time series, a case not covered by generalised estimating equations.

The plan of the paper is as follows. In § 2, a logistic regression model is considered assuming that the data are generated by an nonhomogeneous first-order Markov chain. In § 3, this is extended to the case of missing data and to repeated measures settings. In § 4, some numerical work is presented: specifically, a simulation study has been conducted to investigate some theoretical aspects which could not be studied analytically; furthermore, the methodology is applied to a real data set. Section 5 contains a final discussion.

2. Binary Markov chains

For simplicity, consider first the case of a single stationary process $(Y_1, \ldots, Y_T)$ assumed to be generated by a binary Markov chain taking values 0 and 1. Denote by

$$P = \begin{pmatrix} 1-p_0 & p_0 \\ 1-p_1 & p_1 \end{pmatrix}$$

the transition matrix, where $p_j = \text{pr}\,(Y_t = 1 \mid Y_{t-1} = j)$ for $j = 0, 1$. We search for a parameterisation such that $\theta = E(Y_t)$ is free from the parameter that regulates the serial dependence. A quantity that measures dependence between successive observations is the odds ratio

$$\psi = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{\text{pr}\,(Y_{t-1} = Y_t = 1)\,\text{pr}\,(Y_{t-1} = Y_t = 0)}{\text{pr}\,(Y_{t-1} = 0, Y_t = 1)\,\text{pr}\,(Y_{t-1} = 1, Y_t = 0)}. \tag{1}$$

A technical reason in favour of this choice is given by Fitzmaurice & Laird (1993): when the association between observations is modelled using ψ, the estimates of the mean are relatively insensitive to changes of the association parameter. Moreover, the range of feasible values for ψ is independent of the value of θ.

To obtain (p_0, p_1) for given values of (θ, ψ), we solve (1) and

$$\theta = \theta p_1 + (1-\theta)p_0 \tag{2}$$

with respect to p_0 and p_1.

In practice, we are often concerned with the nonstationary case, in which $\theta_t = E(Y_t)$ varies with t via some function such as

$$\text{logit}\,(\theta_t) = x_t'\beta, \tag{3}$$

where x_t is a k-dimensional vector of time-dependent covariates and β is a k-dimensional parameter. There is no special reason to consider the logit link, which is used only as an example; any other legitimate link function can be used in (3).

In the nonstationary case, we replace (2) by its generalisation

$$\theta_t = \theta_{t-1}p_1 + (1-\theta_{t-1})p_0 \quad (t = 2, \ldots, T), \tag{4}$$

where p_0 and p_1 now vary with t.

For a given value of β, the sequence $\theta_1, \ldots, \theta_T$ is determined by (3), and we can solve (1) and (4) with respect to the p_j's for any $t > 1$. After some algebraic manipulation, we obtain

$$p_j = \begin{cases} \theta_t & \text{for } \psi = 1, \\ \dfrac{\delta - 1 + (\psi-1)(\theta_t - \theta_{t-1})}{2(\psi-1)(1-\theta_{t-1})} + j\,\dfrac{1-\delta+(\psi-1)(\theta_t + \theta_{t-1} - 2\theta_t\theta_{t-1})}{2(\psi-1)\theta_t(1-\theta_{t-1})} & \text{for } \psi \neq 1, \end{cases} \tag{5}$$

$(t = 2, \ldots, T)$, where

$$\delta^2 = 1 + (\psi - 1)\{(\theta_t - \theta_{t-1})^2\psi - (\theta_t + \theta_{t-1})^2 + 2(\theta_t + \theta_{t-1})\}. \qquad (6)$$

It can be shown that the p_j's always lie in $(0, 1)$.

The above relationships allow us to generate a process having the desired properties. On taking $\mathrm{pr}(Y_1 = 1) = \theta_1$ and then generating $Y_2, \ldots, Y_T$ via a nonhomogeneous Markov chain with transition probabilities given by (5), we obtain a sequence such that $E(Y_t) = \theta_t$ for $t = 1, \ldots, T$ and the odds ratios for (Y_{t-1}, Y_t) are equal to ψ.

It is not necessary for ψ to be constant across time. The present formulation concentrates on modelling the mean value θ_t via covariates. However, (5) would apply even if ψ were modelled in terms of covariates, similarly to (3).

Suppose now that a sequence of observed data $y_1, \ldots, y_T$ is available for inference. The log-likelihood function for β and $\lambda = \log \psi$ is

$$l(\beta, \lambda) = \sum_{t=1}^{T} l_t(\beta, \lambda) = \sum_{t=1}^{T} \{y_t \operatorname{logit}(p_{y_{t-1}}) + \log(1 - p_{y_{t-1}})\}, \qquad (7)$$

with p_j defined by (5) for the '$t > 1$' terms of the summation and $p_{y_0} = \theta_1$.

Obtaining the derivatives of $l(\beta, \lambda)$ is conceptually simple but tedious; the relevant expressions are given in the Appendix. It does not seem feasible to obtain expressions for the second order derivatives. Therefore, standard errors of the estimates must be obtained by numerical differentiation of the first derivatives, except in the case of repeated measures where a simpler solution is available, as explained later.

One relevant issue is the orthogonality of β and λ. In the stationary case with $\theta_t \equiv \theta$, one can prove orthogonality of β and λ from the argument of Cox (1970, pp. 72–3), provided end-effects related to the value y_1 and y_T are ignored. Specifically, consider first the case of a single explanatory variable taking constant value. Then (β, λ) form the so-called mixed parameterisation of an exponential family; hence they are orthogonal: see for instance Barndorff-Nielsen & Cox (1994, p. 64). The orthogonality property still holds if θ is a function of a k-dimensional parameter β. It is natural to conjecture that orthogonality also holds in the nonstationary case, but this property has not been proved.

3. Some extensions

3·1. *Missing data*

Consider now the case that some of the observations are missing, but retain the assumption that the designed observation times are equally spaced. In the terminology of Little & Rubin (1987), we assume that the data are missing at random; this means that the reason data are missing at certain time points is independent of the values taken on by the process at those time points.

It is not difficult to generalise (7) to cover this case, since it suffices to replace the one-step conditional probabilities by m-step transition probabilities. If the observations between time t and $t - m$ are missing, we use an expression similar to (7), replacing $p_{y_{t-1}}$ by $p^{(m)}_{t,y_{t-m}}$, where $p^{(m)}_{t,j} = \mathrm{pr}(Y_t = 1 \mid Y_{t-m} = j)$, and summing over the values of the t index corresponding only to observed data. We obtain $p^{(m)}_{t,j}$ by multiplying one-step transition matrices. Again, computing the derivatives of the log-likelihood produces complicated expressions, which are presented in the Appendix.

3·2. *Repeated measures*

An important field of application for the above results is repeated measures, since dependence between successive observations on the same individual must be taken into account, and it is plausible that adjacent data are more strongly correlated than data far apart in time, a feature reproduced by the model under consideration.

Suppose n individuals are available and invididual i is observed at times $1, 2, \ldots, T_i$ $(i = 1, \ldots, n)$ except for some possibly missing data. Denote by y_{it} the observation from subject i at time t, by θ_{it} its expected value, and by x_{it} a k-dimensional vector of covariates associated with design point (i, t).

Assuming that each individual follows the model described in § 2, namely $\operatorname{logit}(\theta_{it}) = x_{it}'\beta$, and that distinct individuals behave independently, the log-likelihood is

$$l(\beta, \lambda) = \sum_{i=1}^{n} l_{i+}(\beta, \lambda),$$

where l_{i+}, is computed applying (7), or its generalisation in the case of missing data, to the sequence $y_{i1}, \ldots, y_{iT_i}$. Similarly, the derivatives of the log-likelihood are obtained by summing the n derivatives computed using the formulae for the case of a single binary time series.

In the context of repeated measures, standard errors of the estimates can be obtained without computing the second derivatives of $l(\beta, \lambda)$. We approximate the variance of the estimates by

$$v_1(\hat\beta, \hat\lambda) = \left[\sum_{i=1}^{n} \begin{pmatrix} \dfrac{\partial l_{i+}}{\partial \beta} \\ \dfrac{\partial l_{i+}}{\partial \lambda} \end{pmatrix} \begin{pmatrix} \dfrac{\partial l_{i+}}{\partial \beta} \\ \dfrac{\partial l_{i+}}{\partial \lambda} \end{pmatrix}'\right]^{-1} \Bigg|_{\beta=\hat\beta, \lambda=\hat\lambda};$$

this approximation is motivated by the fact that the quantity inside square brackets approximates the Fisher information, at least for large n.

4. Some numerical work

4·1. *A simulation study*

A simulation study was conducted with three aims: (i) to explore the conjecture that β and λ are orthogonal or nearly orthogonal parameters in the nonstationary case, (ii) to examine the appropriateness of $v_1(\hat\beta, \hat\lambda)$ in approximating the actual variance matrix of the parameters, and (iii) to test the behaviour of the estimates and their standard errors when the process is erroneously assumed to have constant odds ratio between adjacent time points.

All data sets were generated with $n = 25$, $T_i \equiv 5$, $k = 2$ and the covariates constantly equal to

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -0{\cdot}5 & -0{\cdot}25 & 0 & 0{\cdot}25 & 0{\cdot}5 \end{pmatrix}'$$

for all values of i. For obtaining uniform pseudo-random variates, the algorithm of Wichmann & Hill (1982) was used.

For each data set, the estimate $(\hat\beta_0, \hat\beta_1, \hat\lambda)$ and $v_1(\hat\beta, \hat\lambda)$ were computed. After replicating the above procedure 2500 times, the sample average and sample variance matrix of the 2500 estimated vectors $(\hat\beta_0, \hat\beta_1, \hat\lambda)$ were obtained; we denote by $V_0(\hat\beta, \hat\lambda)$ the sample variance of the estimates. Moreover, the average, $V_1(\hat\beta, \hat\lambda)$, of the 2500 matrices $v_1(\hat\beta, \hat\lambda)$ was computed.

The first group of simulations generated data from the assumed dependence model, i.e. a Markov chain with constant ψ. Table 1 contains a summary of these simulations. The three main conclusions are: (i) the bias of $(\hat\beta_0, \hat\beta_1, \hat\lambda)$ is very small; (ii) the estimated correlations of the parameters support the conjecture of orthogonality between $\hat\lambda$ and $\hat\beta_j$; (iii) the use of $v_1(\hat\beta, \hat\lambda)$ to estimate the variance matrix of the estimates is extremely effective, as the ratios in the last three columns of Table 1 are very close to 1.

The only case for concern is the final one, with parameter $(1, 0, 0)$. Here a single data set with $\hat\lambda \simeq -21$ produced an absurd matrix $v_1(\hat\beta, \hat\lambda)$; in particular, this affected the average of the $\hat\lambda$'s and the last ratio of the diagonal elements V_1 over V_0. In spite of the fact that all of this was due to a

Table 1. *Summary results of simulations under correct model specification*

			Mean			St. deviation			Correlation			Diagonal $(V_1/V_0)^{‡}$		
β_0	β_1	λ	$\hat\beta_0$	$\hat\beta_1$	$\hat\lambda$	$\hat\beta_0$	$\hat\beta_1$	$\hat\lambda$	$(\hat\beta_0, \hat\beta_1)$	$(\hat\beta_0, \hat\lambda)$	$(\hat\beta_1, \hat\lambda)$	β_0	β_1	λ
0	0	0	0·001	0·014	−0·044	0·186	0·530	0·417	0·015	0·001	0·018	1·034	1·028	1·081
		1	0·002	0·011	0·982	0·226	0·579	0·430	0·008	−0·028	0·011	1·022	1·003	1·754
		−1	0·002	0·010	−1·086	0·149	0·463	0·434	0·000	−0·001	−0·003	1·055	1·042	1·088
		2	−0·003	0·004	2·001	0·267	0·576	0·473	−0·015	−0·022	0·007	1·008	1·004	1·057
		−2	0·000	0·005	−2·128	0·119	0·390	0·483	0·011	−0·010	0·020	1·066	1·065	1·084
0	0·5	0	0·001	0·522	−0·045	0·186	0·532	0·421	0·024	−0·008	0·006	1·043	1·034	1·084
		1	0·004	0·524	0·983	0·225	0·578	0·430	0·014	−0·040	−0·001	1·030	1·008	1·082
		−1	0·004	0·520	−1·091	0·151	0·470	0·440	0·002	−0·009	−0·016	1·050	1·042	1·088
0	1	0	0·001	1·032	−0·045	0·187	0·537	0·430	0·000	0·000	−0·023	1·052	1·043	1·083
		1	0·004	1·042	0·978	0·228	0·581	0·434	0·041	−0·027	−0·017	1·029	1·024	1·092
		−1	0·003	1·033	−1·090	0·154	0·475	0·452	0·004	−0·007	−0·023	1·048	1·054	1·091
0·5	0·5	0	0·512	0·526	−0·061	0·194	0·547	0·451	0·006	−0·004	−0·005	1·033	1·039	1·095
		1	0·512	0·527	0·966	0·232	0·592	0·454	0·045	−0·019	0·009	1·031	1·020	1·084
		−1	0·508	0·524	−1·105	0·157	0·489	0·496	0·068	−0·037	0·021	1·065	1·050	1·094
1	0	0	1·021	0·019	−0·116	0·209	0·606	0·689	0·006	−0·070	0·000	1·043	1·041	2×10^8

Table 2. *Summary results of simulations when dependence is incorrectly specified*

			Mean			St. deviation			Correlation			Diagonal $(V_1/V_0)^{‡}$		
β_0	β_1	λ	$\hat\beta_0$	$\hat\beta_1$	$\hat\lambda$	$\hat\beta_0$	$\hat\beta_1$	$\hat\lambda$	$(\hat\beta_0, \hat\beta_1)$	$(\hat\beta_0, \hat\beta_1)$	$(\hat\beta_1, \hat\lambda)$	β_0	β_1	λ
0	0	1	0·000	0·007	−0·062	0·213	0·510	0·509	0·001	−0·007	−0·007	0·774	1·114	0·731
		2	−0·002	0·005	−0·072	0·240	0·449	0·596	0·000	0·000	−0·001	0·611	1·403	0·526
		−1	0·000	0·012	−0·042	0·159	0·521	0·342	0·000	−0·001	0·000	1·437	1·054	1·621
		−2	0·000	0·011	−0·043	0·136	0·494	0·284	0·008	−0·007	0·001	1·986	1·169	2·386
0·5	0·5	1	0·511	0·526	−0·077	0·220	0·527	0·546	0·007	−0·003	−0·003	0·787	1·124	0·760
		−1	0·509	0·523	−0·055	0·166	0·549	0·379	0·006	−0·002	−0·005	1·404	1·035	1·537
0	1	1	0·003	1·040	−0·074	0·215	0·516	0·521	0·001	0·000	−0·004	0·784	1·131	0·737
		−1	0·002	1·036	−0·044	0·160	0·532	0·349	0·021	−0·016	0·004	1·447	1·056	1·636

single case out of 2500, the author felt it inappropriate to downweight the outlying observation using robust methods and decided to summarise the results by plain averages.

In a second group of simulations, the y_{it}'s were generated using the same model for the mean values θ_{it} but changing the dependence structure. The purpose was to examine the effect of incorrect specification of the serial dependence structure on the estimates of the regression parameters, which are usually the parameters of interest. We generated y_{i1} and y_{i2} independently with mean value θ_1 and θ_2, respectively, and the remaining values (y_{i3}, y_{i4}, y_{i5}) by imposing on $(y_{i,t-2}, y_{i,t})$ the kind of dependence which so far we imposed on $(y_{i,t-1}, y_{i,t})$. Table 2 reports a summary of the simulations; these figures indicate a satisfactory degree of robustness of the proposed models, insofar as the mean values of $(\hat\beta_0, \hat\beta_1)$ are almost exactly equal to the theoretical values, near orthogonality of the parameters is preserved and the inaccuracy of estimated standard errors, as measured by the square roots of the ratios V_1/V_0, remains within acceptable bounds, except possibly when $|\lambda|$ is large.

4·2. *A real-data example*

Fitzmaurice, Laird & Lipsitz (1994) have analysed a subset of data from the Muscatine Coronary Risk Factor Study, a longitudinal study of coronary risk factors in school children from Muscatine, Iowa. The data set contains records on 1014 children who were 7–9 years old in 1977 and were examined in 1977, 1979 and 1981. Height and weight were measured in each survey year and those children with relative weight greater than 110% of the median weight in their respective stratum were classified as obese.

The binary response of interest is whether the child is obese (1) or not (0). Since one of the objectives of study was to determine the effects of sex and age on risk of obesity, the marginal probability of obesity for each given value of the risk factors is the quantity of interest. Therefore, a marginal model is appropriate.

However, many data records are incomplete, since many children participated only in one or two occasions of the survey. In particular some of them participated in the 1977 and 1981 surveys, but not in 1979, therefore creating a 'genuine' missing data problem; those subjects having missing observations only in 1977 or 1981 could be regarded as complete but unbalanced data. Fitzmaurice et al. (1994) give a list of the data.

For comparison with the results of Fitzmaurice, Laird & Lipsitz, the same three models for the marginal probability of the event have been fitted to the data, namely:

Model I: $\text{logit}\,(\theta) = \beta_0 + \beta_1 G + \beta_2 A(L) + \beta_3 A(Q) + \beta_4 GA(L) + \beta_5 GA(Q)$,
Model II: $\text{logit}\,(\theta) = \beta_0 + \beta_1 G + \beta_2 A(L) + \beta_3 A(Q)$,
Model III: $\text{logit}\,(\theta) = \beta_0 + \beta_1 A(L) + \beta_2 A(Q)$,

where G indicates sex (female = 1, male = 0) and $A(L)$, $A(Q)$ are orthogonal polynomial contrasts for age effect.

The estimated values of the parameters and their standard errors are reported in Table 3; compare with values given by Fitzmaurice et al. (1994). The parameter estimates are generally quite similar between the two methods. The main differences occur between the standard errors, which are far smaller for the present method. This fact can be explained by the smaller number of nuisance parameters; the present method uses one parameter instead of four to allow for dependence between adjacent observations.

Computer programs in S-PLUS and FORTRAN for the Muscatine data analysis and for the simulation study are available from the author on request.

5. DISCUSSION

Fitzmaurice & Laird (1993) tackled a similar problem to ours, via the so-called 'mixed parameter' model. Their approach offers some desirable features and some disadvantages. Among the advantages are generality and robustness to incorrect modelling of serial dependence. The main disadvantage of the 'mixed parameter' model, as discussed by Fitzmaurice, Laird & Rotmitsky (1993), is

Table 3. *Parameter estimates for Muscatine data*

Model	log *L*	Parameter	Estimate	SE	Ratio
I	−966·56	Intercept	−1·386	0·031	−45·21
		G	0·228	0·040	5·71
		$A(L)$	0·141	0·027	5·30
		$A(Q)$	0·060	0·031	1·91
		$G \times A(L)$	0·160	0·037	4·27
		$G \times A(Q)$	−0·292	0·042	−6·98
		λ	3·066	0·965	3·18
II	−969·52	Intercept	−1·288	0·023	−55·57
		G	0·038	0·022	1·72
		$A(L)$	0·222	0·019	11·64
		$A(Q)$	−0·090	0·020	−4·46
		λ	3·026	0·043	70·42
III	−969·56	Intercept	−1·269	0·020	−63·87
		$A(L)$	0·220	0·019	11·62
		$A(Q)$	−0·090	0·020	−4·43
		λ	3·026	0·043	70·46

that the distribution is not 'reproducible', which means that a subset of length T^* of a series of length T has a distribution different from that obtained by considering only the corresponding subset of observations and parameters. This aspect makes the 'mixed parameter' model inappropriate for analysing series of different lengths. A second disadvantage is that the association parameters are odds ratios of the distribution of the adjacent variables conditional on the remaining data, instead of the more familiar marginal odds ratio. Finally, the number of nuisance parameters grows rapidly when T increases, with possible loss of accuracy of the estimates of interest.

None of these problems are present in the formulation of this paper. Instead, a plausible criticism of this approach could be that the form of serial dependence is quite restricted. When the interest of the analysis focuses on the regression parameters, it seems to us acceptable if serial dependence is not accurately modelled, provided the regression parameters are not strongly influenced. Some evidence in favour of the robustness of the approach is provided by a small simulation study.

Acknowledgement

An earlier version of this paper has greatly benefited from comments and constructive criticism of an associate Editor and two referees. Many thanks are also due to David Cox, Tom DiCiccio and Alessandra Salvan for additional remarks. This work has been supported by the Italian Ministry for University and Research.

Appendix

Derivatives of the log-likelihood

We give expressions to compute the derivatives $\partial l/\partial\beta$ and $\partial l/\partial\lambda$. Assume first that a complete sequence $y_1, \ldots, y_T$ with no missing values is available.

Consider the tth term l_t of the summation in (7); its derivatives are computed via the chain rule, giving

$$\frac{\partial l_t}{\partial \beta} = \frac{\partial l_t}{\partial p_{y_{t-1}}}\left(\frac{\partial p_{y_{t-1}}}{\partial \theta_t}\frac{\partial \theta_t}{\partial \beta} + \frac{\partial p_{y_{t-1}}}{\partial \theta_{t-1}}\frac{\partial \theta_{t-1}}{\partial \beta}\right) \quad (t = 1, \ldots, T),$$

$$\frac{\partial l_t}{\partial \lambda} = \frac{\partial l_t}{\partial p_{y_{t-1}}}\frac{\partial p_{y_{t-1}}}{\partial \psi}\frac{\partial \psi}{\partial \lambda} \quad (t = 2, \ldots, T).$$

These quantities depend in turn on

$$\frac{\partial p_{y_{t-1}}}{\partial \theta_t} = \frac{1}{A}\left\{-2(y_{t-1}-1)\frac{\partial \delta}{\partial \theta_t} + \psi - 1\right\}, \quad \frac{\partial \theta_t}{\partial \beta} = \theta_t(1-\theta_t)x_t,$$

$$\frac{\partial p_{y_{t-1}}}{\partial \theta_{t-1}} = \frac{1}{A^2}\left\{(2y_{t-1}-1)\left(\psi - 1 - \frac{\partial \delta}{\partial \theta_{t-1}}\right)A - 2(\psi-1)(2y_{t-1}-1)B\right\},$$

$$\frac{\partial \theta_{t-1}}{\partial \beta} = \theta_{t-1}(1-\theta_{t-1})x_{t-1},$$

$$\frac{\partial p_{y_{t-1}}}{\partial \psi} = \frac{1}{A^2}\left[\left\{(2y_{t-1}-1)\left(-\frac{\partial \delta}{\partial \psi} + \theta_{t-1}\right) + \theta_t\right\}A - 2B\{1 - y_{t-1} + (2y_{t-1}-1)\theta_{t-1}\}\right],$$

where δ is defined in (6), A and B are the denominator and numerator of $p_{y_{t-1}}$, writing

$$p_j = \frac{B}{A} = \frac{(2j-1)\{1-\delta+(\psi-1)\theta_{t-1}\}+(\psi-1)\theta_t}{2(\psi-1)\{1-j+(2j-1)\theta_{t-1}\}},$$

which is equivalent to (5) when $\psi \neq 1$, and

$$\frac{\partial \delta}{\partial \theta_t} = \frac{1}{\delta}\left[(\psi-1)\{\psi(\theta_t-\theta_{t-1})-(\theta_t+\theta_{t-1})+1\}\right],$$

$$\frac{\partial \delta}{\partial \theta_{t-1}} = \frac{1}{\delta}\left[(\psi-1)\{-\psi(\theta_t-\theta_{t-1})-(\theta_t+\theta_{t-1})+1\}\right],$$

$$\frac{\partial \delta}{\partial \psi} = \frac{1}{2\delta}\{(\theta_t-\theta_{t-1})^2(2\psi-1)-(\theta_t+\theta_{t-1})^2+2(\theta_t+\theta_{t-1})\}.$$

Finally, adding $\partial\psi/\partial\lambda = \psi$, we have all ingredients to compute $(\partial l/\partial\beta, \partial l/\partial\psi)$. If the link function (3) was changed, one would need to change $(\partial\theta_t/\partial\beta, \partial\theta_{t-1}/\partial\beta)$ accordingly.

Consider now the case of missing observations, discussed in § 3·1. In practice, the problem is reduced to the computation of the derivatives of $p_{t,j}^{(m)}$. From the Chapman–Kolmogorov identity

$$p_{t,j}^{(m)} = (1-p_{t-1,j}^{(m-1)})p_{t,0} + p_{t-1,j}^{(m-1)}p_{t,1},$$

we can write

$$\frac{\partial p_{t,j}^{(m)}}{\partial \omega} = -\frac{\partial p_{t-1,j}^{(m-1)}}{\partial \omega}p_{t,0} + (1-p_{t-1,j}^{(m-1)})\frac{\partial p_{t,0}}{\partial \omega} + \frac{\partial p_{t-1,j}^{(m-1)}}{\partial \omega}p_{t,1} + p_{t-1,j}^{(m-1)}\frac{\partial p_{t,j}}{\partial \omega} \qquad \text{(A1)}$$

for a generic quantity ω. This formula must then be used recursively on $p_{t-1,j}^{(m-1)}$ to obtain the full expression of the derivative.

We illustrate the use of (A1) in the case of $m=2$. Without loss of generality, set $t=3$ and consider, with a temporary change of notation, the derivatives of

$$p_{13|j} = \mathrm{pr}\,(Y_3 = 1 \mid Y_1 = j)$$

with respect to θ_t, for $t = 1, 2, 3$. From (A1), we write

$$\frac{\partial p_{13|j}}{\partial \theta_t} = -\frac{\partial p_{12|j}}{\partial \theta_t}p_{23|0} + (1-p_{12|j})\frac{\partial p_{23|0}}{\partial \theta_t} + \frac{\partial p_{12|j}}{\partial \theta_t}p_{23|1} + p_{12|j}\frac{\partial p_{23|1}}{\partial \theta_t}$$

for $t = 1, 2, 3$; a similar expression holds for $\partial p_{13|j}/\partial\psi$. Then we have

$$\frac{\partial p_{13|j}}{\partial\theta_1} = -\frac{\partial p_{12|j}}{\partial\theta_1} p_{23|0} + \frac{\partial p_{12|j}}{\partial\theta_1} p_{23|1},$$

$$\frac{\partial p_{13|j}}{\partial\theta_2} = -\frac{\partial p_{12|j}}{\partial\theta_2} p_{23|0} + (1 - p_{12|j})\frac{\partial p_{23|0}}{\partial\theta_2} + \frac{\partial p_{12|j}}{\partial\theta_2} p_{23|1} + \frac{\partial p_{23|1}}{\partial\theta_2} p_{12|j},$$

$$\frac{\partial p_{13|j}}{\partial\theta_3} = (1 - p_{12|j})\frac{\partial p_{23|0}}{\partial\theta_3} + p_{12|j}\frac{\partial p_{23|1}}{\partial\theta_3}.$$

The derivative $\partial p_{13|j}/\partial\psi$ has a similar pattern of $\partial p_{13|j}/\partial\theta_2$. Finally, from

$$\frac{\partial p^{(2)}_{3,j}}{\partial\beta} = \sum_{t=1}^{3} \frac{\partial p^{(2)}_{3,j}}{\partial\theta_t}\frac{\partial\theta_t}{\partial\beta}$$

we obtain $\partial l/\partial\beta$, and a similar computation produces $\partial l/\partial\lambda$.

References

Barndorff-Nielsen, O. E. & Cox, D. R. (1994). *Inference and Asymptotics*. London: Chapman & Hall.

Carey, V., Zeger, S. L. & Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regression. *Biometrika* **80**, 517–26.

Cox, D. R. (1970). *Analysis of Binary Data*. London: Chapman & Hall.

Cox, D. R. & Snell, E. J. (1989). *Analysis of Binary Data*, 2nd ed. London: Chapman & Hall.

Fitzmaurice, G. M. & Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**, 141–51.

Fitzmaurice, G. M., Laird, N. M. & Lipsitz, S. R. (1994). Analysing incomplete longitudinal responses: a likelihood-based approach. *Biometrics*. To appear.

Fitzmaurice, G. M., Laird, N. M. & Rotnitsky (1993). Regression models for discrete longitudinal data (with discussion). *Statist. Sci.* **8**, 284–309.

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

Lipsitz, S. R., Laird, N. M. & Harrington, D. P. (1992). A three-stage estimator for studies with repeated and possibly missing binary outcomes. *Appl. Statist.* **41**, 203–13.

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

Stiratelli, R., Laird, N. & Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics* **40**, 961–71.

Ware, J. H., Lipsitz, S. & Speizer, F. E. (1988). Issues in the analysis of repeated categorical outcomes. *Statist. Med.* **7**, 95–108.

Wichmann, B. A. & Hill, I. D. (1982). Algorithm AS 183: An efficient and portable pseudo-random number generator. *Appl. Statist.* **31**, 188–90.

Zeger, S. L. & Qaqish, B. (1988). Markov regression models for time series. *Biometrics* **44**, 1019–31.

[*Received March* 1993. *Revised June* 1994]