EXTRACTA MATHEMATICAE 1, n.3, 115-117 (1986)

rth ORDER MARKOV MODELS BASED IN MULTIPLE LOGISTIC REGRESSION FOR COVARIATE DEPENDENCE OF SEQUENCES

Manuel Molina Fernández

Cátedra de Bioestadística. Facultad de Medicina.
Universidad de Extremadura. 06071-Badajoz.

1980 Mathematics Subject Classification : 60J10 , 62M02 , 62M05

A classical problem in many experimental investigations is to model the response variable, subject to observation, in function of the covariates that have influence in the study. The binary case has received a considerable attention (Korn and Whittemore (1979), Muenz and Rubinstein (1985), Pregibon (1982)). In this work, multiple response (or state) is considered and models based in multiple logistic regression are proposed. The basic assumption is that the sequence of states is a rth order Markov chain with stationary transition probabilities. The regression coefficients are estimated by maximum likelihood procedure, and in order to test the order of the chain, a likelihood ratio test is obtained. When r=1, theses problems have been studied, considering several situations, in Molina and Requena (1986), Molina (1986a), and Molina (1986b).

Models based in multiple logistic regression.

Suppose that a heterogeneous group of individuals is followed over time, with each individual passing through states according to a rth order Markov chain with state space $\{1,2,\dots,m\}$ and stationary transition probabilities. Suppose that the observations are made at times $t=0,1,\dots,T$, where T must be at least r, and let $Z_1,\dots,Z_k$ be the covariates.

In order to explain how the covariates relate to changes in state, we proposed to model $p_{q,i_1\dots i_r} = P(X_t=q \,/\, X_{t-r}=i_r,\dots,X_{t-1}=i_1)$, with $q=1,\dots,m-1$, $i_1,\dots,i_r=1,\dots,m$, $t=r,\dots,T$, in the form :

$$p_{q,i_1\dots i_r}=\exp\{f(\boldsymbol{\beta}_{q,i_1\dots i_r};Z_1\dots Z_k)\}(1+\sum_{q=1}^{m-1}\exp\{f(\boldsymbol{\beta}_{q,i_1\dots i_r};Z_1\dots Z_k)\})^{-1}$$

where $f(\boldsymbol{\beta}_{\mathbf{q},\mathbf{i}_1\ldots\mathbf{i}_r};Z_1\ldots Z_k)$ is a linear function of the covariates and the vector of regression coefficients $\boldsymbol{\beta}_{\mathbf{q},\mathbf{i}_1\ldots\mathbf{i}_r}$ whose exact form will vary according to the model under study. For simplicity let us consider the case when $f(\boldsymbol{\beta}_{\mathbf{q},\mathbf{i}_1\ldots\mathbf{i}_r};Z_1\ldots Z_k)=\boldsymbol{\beta}_{\mathbf{q},\mathbf{i}_1\ldots\mathbf{i}_r}.Z'$, where Z is the vector $(1, z_1,\ldots,z_k)$ with z_i the initial value of covariate Z_i, and $\boldsymbol{\beta}_{\mathbf{q},\mathbf{i}_1\ldots\mathbf{i}_r}=(\beta^0_{q,i_1\ldots i_r},\ldots,\beta^k_{q,i_1\ldots i_r})$. Such situation corresponding to model without interaction between covariates.

Estimating the regression coefficients.

Let the sequences observed be $x_{oh},x_{1h},\ldots,x_{Th}$ $(h=1,\ldots,n)$, in n individual up to time T, and let $Z_h=(1,z_{1h},\ldots,z_{kh})$ be the vector of covariates for the hth individual. If the r initials states of each individual are known, then the log-likelihood function can be written in the form :

$$L_T(\boldsymbol{\beta}_{\mathbf{i}_1\ldots\mathbf{i}_r},\ i_1,\ldots,i_r=1,\ldots,m)=\sum_{i_1,\ldots,i_r=1}^{m} F_{i_1\ldots i_r}(\boldsymbol{\beta}_{\mathbf{i}_1\ldots\mathbf{i}_r}) \tag{1}$$

where

$$F_{i_1\ldots i_r}(\boldsymbol{\beta}_{\mathbf{i}_1\ldots\mathbf{i}_r})=\sum_{h=1}^{n}\left(\sum_{q=1}^{m-1} n^h_{q,i_1\ldots i_r}\boldsymbol{\beta}_{\mathbf{q},\mathbf{i}_1\ldots\mathbf{i}_r}.Z'_h - n^h_{.,i_1\ldots i_r}\ln(1+\sum_{q=1}^{m-1}\exp\{\boldsymbol{\beta}_{\mathbf{q},\mathbf{i}_1\ldots\mathbf{i}_r}.Z'_h\})\right)$$

with $\boldsymbol{\beta}_{\mathbf{i}_1\ldots\mathbf{i}_r}=(\boldsymbol{\beta}_{\mathbf{1},\mathbf{i}_1\ldots\mathbf{i}_r}\ldots\boldsymbol{\beta}_{\mathbf{m-1},\mathbf{i}_1\ldots\mathbf{i}_r})$.

$n^h_{q,i_1\ldots i_r}$ is the number of transitions to state q, from the sequence of consecutives states $i_r,i_{r-1},\ldots,i_1$, observed in the hth individual and $n^h_{.,i_1\ldots i_r}=\sum_{q=1}^{m} n^h_{q,i_1\ldots i_r}$

It follow from (1) that for each $i_1,\ldots,i_r\in\{1,\ldots,m\}$, the maximum likelihood estimate cf $\boldsymbol{\beta}_{\mathbf{i}_1\ldots\mathbf{i}_r}$ can be obtained separately. For this purpose, the likelihood equation is :

$$\frac{\partial F_{i_1\ldots i_r}(\boldsymbol{\beta}_{\mathbf{i}_1\ldots\mathbf{i}_r})}{\partial\beta^s_{q,i_1\ldots i_r}}=\sum_{h=1}^{n}(n^h_{q,i_1\ldots i_r}-n^h_{.,i_1\ldots i_r}p^h_{q,i_1\ldots i_r})z_{sh}=0 \tag{2}$$

with $s=o,\ldots,k$, $q=1,\ldots,m-1$, and $z_{oh}=1$ $(h=1,\ldots,n)$

We can maximize (2) by Newton-Raphson method as long as the second derivatives matrix has full rank, and $\min_{i_1,\ldots,i_r} \{ \sum_{h=1}^{n} n^h_{\cdot,i_1\ldots i_r} \} \geq k+1$

For large n, an estimate of the covariance matrix of $\hat{\beta}_{i_1\ldots i_r}$, is the matrix that has elements equal to -1 times the inverse of the second derivatives matrix evaluated in $\hat{\beta}_{i_1\ldots i_r}$.

Test of the hypothesis that the chain is of a given order.

For testing the hypothesis that the process is a uth order Markov chain against the alternative it is rth but not uth order $(u < r)$, the statistic $-2(L_T(\hat{\beta}_{i_1\ldots i_u}) - L_T(\hat{\beta}_{i_1\ldots i_r}))$, that is asymptotically distributed as χ^2 with $(m-1)m^u(m^{r-u}-1)(k+1)$ degree of freedom, can be used.

References.

Korn, E.L. and Whittemore, A.S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. Biometrics, 35 , 795-802.

Molina, M. (1986a). Modelos markovianos dependientes de covariables para el estudio de sucesiones de estados no necesariamente binarios. XI Jornadas Hispano-Lusas de Matemáticas. Badajoz.

Molina, M. (1986b). Modelos markovianos no estacionarios para el estudio de secuencias no necesariamente binarias dependientes de covariables. I Conferencia Española de Biometria. Granada.

Molina, M. y Requena, F. (1986). Sobre la estimación de maxima verosimilitud de los coeficientes de regresión logística en modelos markovianos con dependencia de covariables. C. de Bioestadística y sus aplicaciones informáticas. (Por aparecer)

Muenz, L.R. and Rubinstein, L.M. (1985). Markov models for covariate dependence of binary sequences. Biometrics, 41, 91-101.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. Biometrics, 38, 485-498.