EXTRACTA MATHEMATICAE Vol. **10**, Núm. 2, 140 – 145 (1995)

A Goodness of Fit Test in Markov Models with Dependence of Covariates

M. Barrantes [1], M. González [2] and M. Molina [2]

[1] *Depto. de Didáctica de Ciencias Experimentales y Matemáticas Universidad de Extremadura,* 06071-*Badajoz, Spain*

[2] *Depto. de Matemáticas, Universidad de Extremadura,* 06071-*Badajoz, Spain*

AMS *Subject Class.* (1991): 60J10, 62M02

Received October 19, 1995

1. Introduction

In this paper we study how to fit the probability distribution of a nominal-scaled response variable in function of the covariates with influence on the response. The binary response case has been amply investigated and specially fitted by logistic regression (Bonney [1], Davison [2], Lilienfield and Pyne [8], Muenz and Rubinstein [12], ...). For this situation, different methods which enable us to evaluate and improve the fit have been developed (Fowlkes [3], Kay and Little [5], Minkin [9], Pregibon [13], Hosmer and Lemeshow [4], ...). Suitable extensions for the multiple response case have been introducted (Lesaffre and Albert [6], Liang [7], Santner and Duffy [14], ...), however diagnostic methods for this situation have not been sufficiently investigated.

In recent paper, Molina and González [10], under the assumption that the response change is well described by a non-stationary r-th order Markov chain, the transition probabilities were modeled through the Multiple-group Logistic Regression Model (MLRM). The maximum likelihood estimation of the regression parameters was considered and a method to evaluate whether the logistic model is the correct one to fit was derived.

In this work, a test is proposed for the purpose of assessing the goodness of fit of the MLRM to transition probabilities in the Markov model. The method suggested is a generalization, to multiple response and adaptated for Markov chains, of the one studied by Hosmer and Lemeshow [4].

2. Mathematical Model

Suppose that a nominal-scaled variable with m possible responses ($2 \leq m < \infty$) and influenced by the covariates $Z_1, \ldots, Z_k$, is observed regularly in time. We assume that the underlying process is an r-th order Markov chain $\{X_n : n = 0, 1, \ldots\}$, ($r \geq 1$), with state space $S = \{1, \ldots, m\}$ and non stationary transition probabilities, where $X_n = i$, if at time n the observed response is the i-th. We consider the vector $G_n = (G_{1n}, \ldots, G_{kn})$, where $G_{in} = G_{in}(Z_{i0}, \ldots, Z_{i(n-1)})$ is a $\mathbb{R}$-valued Borel-measurable function on $\mathbb{R}^n$, Z_{it} being the covariate Z_i observed at time t. In this situation, we model the transition probabilities, namely $Pr[X_n = i \,|\, (X_{n-1}, \ldots, X_{n-r}) = s, G_n = g_n]$, $n \geq r$, $i \in S$, $s \in S^r$, which will be denoted by $p_{si}(n)$, in the form:

$$(2.1)\quad p_{si}(n) = \exp\{\beta^n_{si} \cdot \bar{g}'_n\}\bigl(1 + \sum_{j=1}^{m-1} \exp\{\beta^n_{sj} \cdot \bar{g}'_n\}\bigr)^{-1}, \quad i = 1, \ldots, m-1$$

where $\bar{g}_n = (1, g_n)$ and for $j = 1, \ldots, m-1$, $\beta^n_{sj} = (\beta^n_{sj0}, \ldots, \beta^n_{sjk})$ are parameter vectors.

Thus, each row of the transition matrix is fitted by a different MLRM which has a total number of $(m-1)(k+1)$ parameters to estimate.

Suppose that a sample of N individuals is observed until time T, ($T \geq r$). For $n = 0, \ldots, T$ and $t = 0, \ldots, T-1$, let x^q_n and z^q_{it} be the observed values of the variables X_n and Z_i, at time n and at time t, respectively, for the q-th individual ($q = 1, \ldots, N$). From now on, G_n and $p_{si}(n)$, evaluated on the observations of the q-th individual, will be denoted by g^q_n and $p^q_{si}(n)$, respectively. Then, from (2.1) it is easy to verify that the associated log-likelihood, denoted by L_T, may be written in the form:

$$(2.2)\; L_T = \sum_{s\in S^r} \sum_{n=r}^{T} \sum_{q=1}^{N} \left[\sum_{j=1}^{m-1} \delta^q_{sjn} \beta^n_{sj} \cdot \bar{g}^{q\prime}_n - \delta^q_{sn} \log\bigl(1 + \sum_{l=1}^{m-1} \exp\{\beta^n_{sl} \cdot \bar{g}^{q\prime}_n\}\bigr) \right]$$

being $\bar{g}^q_n = (1, g^q_n)$ and $\delta^q_{sn} = \sum_{j=1}^{m} \delta^q_{sjn}$ with $\delta^q_{sjn} = 1$ if $(x^q_{n-1}, \ldots, x^q_{n-r}) = s$ and $x^q_n = j$, or 0 otherwise.

Really, the full log-likelihood includes a term for the probability of the first r states, but since we want to estimate only the parameters of the transition probabilities, for us this term is non informative.

From (2.2), and for s and n given, it is deduced that the maximum likelihood estimation of β^n_{sj}, $j = 1, \ldots, m-1$, is obtained solving the likelihood equations:

$$\sum_{q=1}^{N} [\delta^q_{sjn} - \delta^q_{sn} p^q_{sj}(n)] g^q_{un} = 0\,, \quad j = 1,\dots,m-1,\ u = 0,\dots,k$$

with $g^q_{0n} = 1$, $q = 1,\dots,N$.

These nonlinear equations must be solved in an iterative manner. The Newton-Raphson procedure can be used, whenever the observed values of the covariates are such that the second partial derivate matrix of the log-likelihood is non-singular (see Molina and González [10]). Let $\hat\beta^n_{sj}$ be the maximum likelihood estimation of β^n_{sj}. We will denote by $\hat p_{sj}(n)$, the transition probability $p_{sj}(n)$ evaluated in $\hat\beta^n_{sj}$.

3. Goodness of Fit Test

For s and n given, we suppose that we wish to test the adequation of the MLRM to the corresponding transition probabilities, i.e., the null hyphotesis will be that $p_{sj}(n)$, $j = 1,\dots,m-1$, are of the form specified in (2.1). For this purpose, the following goodness of fit test, based in the former sample, could be used.

Let $I(s,n) = \{q \in \{1,\dots,N\} : (x^q_{n-1},\dots,x^q_{n-r}) = s\}$ and we denote by $N(s,n)$ the number of the indices falling in $I(s,n)$, (we will assume that N is large enough for that $N(s,n) > 0$, obviously $\sum_{s\in S^r} N(s,n) = N$). We consider the $m-1$ partitions of the interval [0,1]:

$$0 = c^i_0(s,n) < c^i_1(s,n) < \cdots < c^i_{g_i-1}(s,n) < c^i_{g_i}(s,n) = 1\,, \quad i = 1,\dots,m-1$$

being g_i a positive integer ($2 < g_i < \infty$) and we define the $(m-1)$-dimensional random vector W in the form:

For the q-th individual, ($q \in \{1,\dots,N(s,n)\}$)

$$W = (h_1,\dots,h_{m-1})\,,\ h_i = 1,\dots,g_i\,, \qquad \text{if } \hat p^q_{si}(n) \in [c^i_{h_i-1}(s,n), c^i_{h_i}(s,n)) \quad i = 1,\dots,m-1\,.$$

For $l = 1,\dots,m$, $h_i = 1,\dots,g_i$, we denote by $O_{(l,(h_1,\dots,h_{m-1}))}(s,n)$, the "observed frequency" of the pair $[X_n = l, W = (h_1,\dots,h_{m-1})]$ in the sample corresponding to $N(s,n)$ individuals. Now, under the null hyphotesis, the "expected frequency", namely $E_{(l,(h_1,\dots,h_{m-1}))}(s,n)$, will be:

(3.1) $$E_{(l,(h_1,\dots,h_{m-1}))}(s,n) =$$

$$= N(s,n) \int_{c^1_{h_1-1}(s,n)}^{c^1_{h_1}(s,n)} \cdots \int_{c^{m-1}_{h_{m-1}-1}(s,n)}^{c^{m-1}_{h_{m-1}}(s,n)} q_l(s,n) f(y_1,\ldots,y_{m-1}) dy_1 \ldots dy_{m-1}$$

where

$$q_l(s,n) = \begin{cases} p_{sl}(n) & \text{if } l = 1,\ldots,m-1 \\ 1 - \sum_{j=1}^{m-1} p_{sj}(n) & \text{if } l = m \end{cases}$$

being f the density (or probability) function of $(p_{s1}(n),\ldots,p_{s(m-1)}(n))$ considered as function of the random vector $(G_{1n},\ldots,G_{kn})$.

Taking into account the sample of $N(s,n)$ individuals, a estimation of f, will be:

$$\hat{f}(y_1,\ldots,y_{m-1}) = \tag{3.2}$$

$$\begin{cases} N(s,n)^{-1} & \text{if } (y_1,\ldots,y_{m-1}) \in \{(\hat{p}^q_{s1}(n),\ldots,\hat{p}^q_{s(m-1)}(n)) : q = 1,\ldots,N(s,n)\} \\ 0 & \text{otherwise} \end{cases}$$

Consequently replacing (3.2) in (3.1), we have that:

$$\hat{E}_{(l,(h_1,\ldots,h_{m-1}))}(s,n) =$$

$$\begin{cases} \sum\limits_{q \in J_{(h_1,\ldots,h_{m-1})}(s,n)} \hat{p}^q_{sl}(n) & \text{if } l = 1,\ldots,m-1 \\ N_{(h_1,\ldots,h_{m-1})}(s,n) - \sum\limits_{j=1}^{m-1} \sum\limits_{q \in J_{(h_1,\ldots,h_{m-1})}(s,n)} \hat{p}^q_{sj}(n) & \text{if } l = m \end{cases}$$

where $J_{(h_1,\ldots,h_{m-1})}(s,n) = \{q \in \{1,\ldots,N(s,n)\} : \hat{p}^q_{sl}(n) \in [c^l_{h_l-1}(s,n), c^l_{h_l}(s,n)), l = 1,\ldots,m-1\}$ and $N_{(h_1,\ldots,h_{m-1})}(s,n)$ is the number of the indices falling in $J_{(h_1,\ldots,h_{m-1})}(s,n)$. The goodness of fit test will be derived comparing the observed frequencies with the expected frequencies through the statistic:

$$H(s,n) = \sum_{l=1}^{m} \sum_{h_1=1}^{g_1} \cdots \sum_{h_{m-1}=1}^{g_m} \frac{\left[O_{(l,(h_1,\ldots,h_{m-1}))}(s,n) - \hat{E}_{(l,(h_1,\ldots,h_{m-1}))}(s,n)\right]^2}{\hat{E}_{(l,(h_1,\ldots,h_{m-1}))}(s,n)}.$$

The asymptotic distribution (when $N(s,n) \to \infty$) of $H(s,n)$ can not be obtained from a direct application of the usual theory used for chi-squared goodness of fit tests (mainly, because the observed frequencies are based on the estimations of the parameters β^n_{sj}, $j = 1, \ldots, m-1$). But, having use of the theory for chi-squared test of Moore and Spruill [11], it is deduced that

$$H(s,n) \to \chi^2_{(\nu)} + \sum_{i=1}^{(k+1)(m-1)} \mu_i \cdot \chi^2_{(1)} \qquad \text{when } N(s,n) \to \infty \tag{3.3}$$

where $\nu = m \prod_{i=1}^{m-1} g_i - km + k - m$, and μ_i, $i = 1, \ldots, (k+1)(m-1)$, are the non-zero or 1 eigenvalues of the matrix:

$$Q(s,n) = I(s,n) - U(s,n)' \cdot U(s,n) - V(s,n)J(s,n)^{-1}V(s,n)'$$

where $I(s,n)$ is the $m \prod_{i=1}^{m-1} g_i \times m \prod_{i=1}^{m-1} g_i$ identity matrix,

$$U(s,n) = N(s,n)^{-1}(\hat{E}_1(s,n) \ldots \hat{E}_m(s,n)),$$

being $\hat{E}_l(s,n) = (\hat{E}_{(l,(h_1,\ldots,h_{m-1}))}(s,n), h_i = 1, \ldots, g_i, i = 1, \ldots, m-1)$, $l = 1, \ldots, m$ $V(s,n)$ is the $m \prod_{i=1}^{m-1} g_i \times (k+1)(m-1)$ matrix which has as general element

$$N(s,n)^{1/2}\big(\hat{E}_{(l,(h_1,\ldots,h_{m-1}))}(s,n)\big)^{-1/2}\big(\partial E_{(l,(h_1,\ldots,h_{m-1}))}(s,n)/\partial\beta^n_{sju}\big)$$

$$(l = 1, \ldots, m\,,\ h_i = 1, \ldots, g_i\,,\ j = 1, \ldots, m-1\,,\ u = 0, \ldots, k)$$

and $J(s,n)$ is the $(k+1)(m-1) \times (k+1)(m-1)$ information matrix, (evaluated at the true parameters values). In the practical applications, the component $\sum_{i=1}^{(k+1)(m-1)} \mu_i \chi^2_{(1)}$ in (3.3) is well approximated through a chi-squared distribution.

References

[1] Bonney, G., Logistic regression for dependent binary observations, *Biometrics* **43** (1987), 951–973.

[2] Davison A., Approximate conditional inference in generalized linear models, *J. Royal Statistics, Ser. B* **50** (1988), 445–461.

[3] Fowlkes, E., Some diagnostics for binary logistic regression via smoothing, *Biometrika* **74** (1987), 503–515.

[4] Hosmer, D., Lemeshow, S., Goodness of fit test for the multiple logistic regression model, *Commun. Statist. Theor. Meth.* **A9** (10) (1980), 1043–1069.

[5] Kay, R., Little, S., Transformations of the explanatory variables in the logistic regression model for binary data, *Biometrika* **74** (1987), 495–501.

[6] Lesaffre, E., Albert, A., Multiple-group logistic regression diagnostic, *Applied Statistics* **38** (1989), 425–440.

[7] Liang, K., Extended Mantel-Haenszel estimating procedure for multivariate logistic regression models, *Biometrics* **43** (1987), 289–299.

[8] LILIENFELD, D., PYNE, D., The logistic analysis of epidemiologic prospective studies: investigation by simulation, *Statistics in Medicine* **3** (1984), 15–24.

[9] MINKIN, S., Fit assessment and identification of functional form in logistic regression, *Applied Statistics* **38** (1989), 343–350.

[10] MOLINA, M., GONZÁLEZ, M., Modelization in non stationary finite Markov chain, to appear in *Revista di Matematica Pura ed Applicata* **20** (1995).

[11] MOORE, D., SPRUILL, M., Unified large-sample theory of general chi-squared statistics for test of fit, *Annals of Statistics* **3** (1975), 599–616.

[12] MUENZ, L., RUBINSTEIN, L., Markov models for covariate dependence of binary sequences, *Biometrics* **41** (1985), 91–101.

[13] PREGIBON, D., Logistic regression diagnostic, *Annals of Statistics* **9** (1981), 705–724.

[14] SANTNER, T., DUFFY, D., A note on A. Albert and J. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models, *Biometrika* **73** (1986), 755–758.