International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 3 (Special Issue), 2004, pp 241–249

A Three State Markov Model for Analyzing Covariate Dependence

M. Ataharul Islam
Department of Statistics and Operations Research
King Saud University
Saudi Arabia

Rafiqul Islam Chowdhury
Department of Health Information Administration
Kuwait University
Kuwait

[Received April 27, 2004; Accepted November 5, 2004]

Abstract

This paper examines the covariate dependence in a three state Markov model. Muenz and Rubinstein (1985) introduced a Markov model of first order for analyzing covariate dependence employing logistic regression. In this paper, the covariate dependent Markov model is proposed for more than two states. For the sake of notational convenience, a three state Markov model is considered that can be increased to any finite number of states. The proposed model can be used for a wide range of practical applications. This paper shows the estimation and test procedures for a covariate dependent three state Markov model. The model is illustrated for analyzing longitudinal data on pregnancy complications.

Keywords and Phrases: Markov Model, Covariate Dependence, Logistic Regression, Three States, Higher Order.

AMS Classification: 60J20.

1 Introduction

In analyzing the longitudinal data, the use of Markov chain models have increased to a large extent during the recent past. Following papers indicate the variety of work in the use of Markov models: (i) Regier (1968) introduced a two state transition matrix

for estimating odds ratio, (ii) Prentice and Gloeckler (1978) proposed a grouped data version of the proportional hazards regression model for estimating computationally feasible estimators of the relative risk function, (iii) Korn and Whittemore (1979) proposed a model in order to incorporate role of previous state as a covariate to analyze the probability of occupying the current state, (iv) Muenz and Rubinstein (1985) introduced a discrete time Markov chain for expressing the transition probabilities in terms of function of covariates for a binary sequence of presence or absence of a disease.

Among the more recent works, noteworthy are Albert (1994), Albert and Myron (1998), Raftery and Tavare (1994). In recent years, there is a great deal of interest in the development of multivariate models based on the Markov Chains. These models can be employed for analyzing data generated from meteorology, epidemiology and survival analysis, reliability, econometric analysis, biological concerns, etc. Muenz and Rubinstein (1986) employed logistic regression models to analyze the transition probabilities from one state to another. The technique proposed by Muenz and Rubinstein considers only two intercommunicating states.

In this paper, a Markov chain model for three intercommunicating states has been proposed to analyze the covariate dependence of the transition probabilities. The risk factors that contribute to specific transitions can be identified from the proposed model.

2 Covariate Dependent First Order Model

A brief overview of the covariate dependent first order Markov model proposed by Muenz and Rubinstein (1986) is presented in this section. Let us consider a two state Markov chain for a discrete time binary sequence as follows:

$$\pi = \begin{bmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{bmatrix}$$

We can define the following: $\pi_{00} = 1-\pi_{01}$ and $\pi_{10} = 1-\pi_{11}$. Probability of a transition from 0 at time t_{j-1} to 1 at time t_j is $\pi_{01} = P(Y_j = 1/Y_{j-1} = 0)$ and similarly the probability of a transition from 1 at time t_{j-1} to 1 at time t_j is

$$\pi_{11} = P(Y_j = 1/Y_{j-1} = 1).$$

The transition probabilities can be defined in terms of function of the covariates as follows:

$$\pi_{01}(Y_j = 1/Y_{j-1}0, X) = \frac{e^{\beta_0' X}}{1 + e^{\beta_0' X}}, \tag{1}$$

and

$$\pi_{11}(Y_j = 1/Y_{j-1} = 1, X) = \frac{e^{\beta_1' X}}{1 + e^{\beta_1' X}} \tag{2}$$

where

$$\begin{aligned} X_i' &= [1, X_{i1}, \cdots, X_{ip}] = \text{vector of covariates for the ith person;} \\ \beta_0' &= [\beta_{00}, \beta_{01}, \cdots, \beta_{0p}] = \text{vector of parameters for the transition from 0;} \\ \beta_1' &= [\beta_{10}, \beta_{11}, \cdots, \beta_{1p}] = \text{vector of parameters for the transition from 1.} \end{aligned}$$

Then the likelihood function can be defined as

$$L = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[\{\pi_{00}\}^{\delta_{00ij}} \{\pi_{01}\}^{\delta_{01ij}} \right] \left[\{\pi_{10}\}^{\delta_{10ij}} \{\pi_{11}\}^{\delta_{11ij}} \right] \tag{3}$$

where n_i = total number of follow-up observations since the entry into the study for the ith individual; $\delta_{msij} = 1$ if a transition type m-s is observed during jth follow-up for the ith individual (m,s=0,1), 0 otherwise. The log likelihood function, after substituting (1) and (2) in (3), can be expressed as

$$\ln L = \ln L_0 + \ln L_1$$

where L_0 corresponds to the first part and L_1 corresponds to the second part of (3).

Hence,

$$\ln L_0 = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left[\delta_{01ij} \left\{ \beta_{01}' X_i \right\} - (\delta_{00ij} + \delta_{01ij}) \ln\{1 + e^{\beta_{01}' X_i}\} \right]$$

and

$$\ln L_1 = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left[\delta_{11ij} \left\{ \beta_{11}' X_i \right\} - (\delta_{10ij} + \delta_{11ij}) \ln\{1 + e^{\beta_{11}' X_i}\} \right].$$

Differentiating with respect to the parameters and solving the following equations we obtain the likelihood estimates for 2(p+1) parameters:

$$\frac{\partial \ln L_0}{\partial \beta_{01q}} = 0, \quad q = 1, 2, \cdots, p;$$

and

$$\frac{\partial \ln L_1}{\partial \beta_{11q}} = 0, \quad q = 1, 2, \cdots, p.$$

3 Three State Markov Model

In reality, we have to face more than two states in many different situations. As an example, we may consider three states of health status, normal, moderately sick, and severely sick. Then we can show the Markov Chain as follows:

$$\pi = \begin{bmatrix} \pi_{00} & \pi_{01} & \pi_{02} \\ \pi_{10} & \pi_{11} & \pi_{11} \\ \pi_{20} & \pi_{21} & \pi_{22} \end{bmatrix}$$

where $\pi_{00} = 1 - \pi_{01} - \pi_{02}$, $\pi_{10} = 1 - \pi_{11} - \pi_{12}$, $\pi_{20} = 1 - \pi_{21} - \pi_{22}$ and . Here, 0 and 1 and 2 are the three possible outcomes of a dependent variable, Y. The probability of a transition from m $(m = 0, 1, 2)$ at time t_{j-1} to s $(s = 0, 1, 2)$ at time t_j is $\pi_{ms} = P(Y_j = s/Y_{j-1} = m)$. It is evident that for any m, $\sum_{s=0}^{2} \pi_{ms} = 1$, $m = 0, 1, 2$.

Let us define the following notations: $X_i = [1, X_{i1}, \cdots, X_{ip}]$ = vector of covariates for the ith person; $\beta'_{ms} = [\beta_{ms0}, \beta_{ms1}, \cdots, \beta_{msp}]$= vector of parameters for the transition from m to s.

Then the transition probabilities can be defined as conditional probabilities in terms of function of the covariates as follows (Hosmer and Lemeshow, 1989):

$$\pi_{ms}(Y_j = s/Y_{j-1} = m, X) = \frac{e^{g_{ms}(X)}}{\sum_{k=0}^{2} e^{g_{mk}(X)}}, \quad m = 0, 1, 2 \tag{4}$$

where

$$g_{ms}(X) = \begin{cases} 0, & \text{if s = 0} \\ \ln\left[\frac{\pi_{ms}(Y_j=s/Y_{j-1}=m,X)}{\pi_{ms}(Y_j=0/Y_{j-1}=m,X)}\right] & \text{if s =1,2.} \end{cases}$$

Hence

$$g_{ms}(X) = \beta_{ms0} + \beta_{ms1}X_1 + \cdots + \beta_{msp}X_p.$$

Then the likelihood function for n individuals with each individual having n_i $(i = 1, 2, \cdots, n)$ follow-ups can be expressed as

$$L = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \prod_{m=0}^{2} \prod_{s=0}^{2} \left[\{\pi_{ms}\}^{\delta_{msij}}\right] \tag{5}$$

where n_i = total number of follow-up observations since the entry into the study for the ith individual; $\delta_{msij} = 1$ if a transition type $m \to s$ is observed during jth follow-up for the ith individual, $\delta_{msij} = 0$, otherwise, $m, s = 0, 1, 2$. The log likelihood function, after substituting (1) and (2) in (3), can be expressed as

$$\ln L = \sum_{m=0}^{2} \ln L_m,$$

where L_m corresponds to the m-th component of the likelihood function.

Hence,

$$\ln L_m = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left[\sum_{s=0}^{2} \delta_{msij} g_{ms}(X_i) - \ln\left(\sum_{k=0}^{2} e^{g_{mk}(X_i)}\right)\right]$$

Differentiating with respect to the parameters and solving the following equations we obtain the likelihood estimates for $6(p+1)$ parameters:

$$\frac{\partial \ln L_m}{\partial \beta_{msq}} = \sum_{i=1}^{n}\sum_{j=1}^{n_i} X_{qi}(\delta_{msij} - \pi_{msij}), \quad q = 0, 1, 2, \cdots, p;\ m = 0, 1, 2.$$

The observed information matrix can be obtained from the following second derivatives:

$$\frac{\partial^2 \ln L_m}{\partial \beta_{msq} \partial \beta_{msq'}} = -\sum_{i=1}^{n}\sum_{j=1}^{n_i} X_{q'i} X_{qi} \pi_{msij}(1 - \pi_{msij}),$$

$q, q' = 0, 1, 2, \cdots, p;\ s = 0, 1, 2;\ m = 0, 1, 2,$ and

$$\frac{\partial^2 \ln L_m}{\partial \beta_{msq} \partial \beta_{ms'q'}} = -\sum_{i=1}^{n}\sum_{j=1}^{n_i} X_{q'i} X_{qi} \pi_{msij} \pi_{ms'ij},$$

$q, q' = 0, 1, 2, \cdots, p;\ s, s' = 0, 1, 2;\ m = 0, 1, 2.$

4 Test of Hypothesis

A straightforward test procedure was first proposed by Anderson and Goodman (1957), Billingsley (1961), and then used by Kalbfleisch and Lawless (1985) and Raftery and Tavarey (1994). The modified test statistic for the multistate Markov model is defined as follows assuming equal number of follow-ups for each subject ($n_i = r$).

$$\chi^2 = \sum_{m=0}^{2}\sum_{s=0}^{2}\sum_{i=1}^{n}\sum_{j=1}^{r} \frac{\{n_{msij} - e_{msij}\}^2}{e_{msij}}$$

where essentially $n_{msij} = \delta_{msij}$ and e_{msij} is the expected number corresponding to the observed number of transitions n_{msij}. The expected number of transitions, n_{msij}, can be obtained from the following steps:

(i) estimate for π_{msij} needs to be obtained for given values of X_{msi} which is $\hat{\pi}_{msij}$; and

(ii) then we have $e_{msij} = (\hat{\pi}_{msij}).(n_{msi.})$ where $n_{msi.} = \sum_{j=1}^{n_i} n_{msij}$.

If none of the transition probabilities is restricted to be zero then the test statistic is the familiar Pearson statistic with $9n(r-1)$ degrees of freedom. The degrees of freedom will be further reduced for the estimates of transition probabilities zero or nearly zero (Kalbfleisch and Lawless, 1985).

The vectors of 6 sets of parameters for the three state Markov model can be represented by the following vector:

$$\beta = [\beta_1, \beta_2, \cdots, \beta_6].$$

To test the null hypothesis $H_0 : \beta = 0$, we can employ the usual likelihood ratio test

$$-2[\ln L(\beta_0) - \ln L(\beta)] \approx \chi^2_{6p}.$$

To test the significance of the qth parameter of the 6 sets of parameters, the null hypothesis is $H_0 : \beta_{mq} = 0$ and the corresponding Wald test is

$$W = \frac{\hat{\beta}_{mq}}{se(\hat{\beta}_{mq})}.$$

5 Application

This study employs data from the survey on Maternal Morbidity in Bangladesh conducted by the Bangladesh Institute for Research for Promotion of Essential and Reproductive Health Technologies (BIRPERHT) during November 1992 to December 1993. The data were collected using both cross-sectional and prospective study designs. This study is based on the data from the prospective component of the survey. The subjects comprised of pregnant women with less than 6 months duration. All the selected pregnant women were followed on regular basis (roughly at an interval of one month) throughout the pregnancy. During the follow-up visits, pregnancy complications were recorded. A total of 1020 pregnant women were interviewed in the follow-up component of the study. For the purpose of this study, we have selected 993 pregnant women, with at least one antenatal follow-up. The following pregnancy complications are considered under the complications in this study: hemorrhage, fits, convulsion, edema, excessive vomiting, and cough or fever for more than three days. If one or more of hemorrhage, fits, convulsion occurred to the respondents, we considered as major complications and coded was as 2, if edema, excessive vomiting, and cough or fever for more than three days occurred to the respondents, we considered as minor complications and was coded as 1, if no complications then coded as 0.

The explanatory variables are: age at marriage (15 years or lower, more than 15 years), index pregnancy was wanted or not (no, yes), visit to health care facilities (yes, no), pregnancies prior to the index pregnancy (yes, no), and education of respondent (no schooling, some schooling).

The number of transitions for the three-state Markov chain of first order is displayed in Table 1. The estimates of parameters of covariate dependent Markov models are presented in Table 2.

Table 2 provides estimates from some of the transitions for a three-state Markov model with covariate dependence for analyzing the pregnancy complications. The

Table 1: Number of Transitions for Pregnancy Complications

Transitions	$\rightarrow 0$	$\rightarrow 1$	$\rightarrow 2$
First Order			
$0\rightarrow$	1416	275	17
$1\rightarrow$	363	602	31
$2\rightarrow$	41	39	50

Table 2: Estimates of Parameters of Covariate Dependent Markov Models for Analyzing Pregnancy Complications

Variables	**Estimates**	**Std. error**	**t-value**	**p-value**
$0\rightarrow$ 1				
Constant	-1.31	0.155	-8.44	0.000
Age at marriage ($\leq 14 = 1$)	-0.06	0.099	-0.59	0.553
Wanted pregnancy (Yes=1)	-0.19	0.107	-1.76	0.078
Visit to HC (yes=1)	0.61	0.107	5.69	0.000
Previous pregnancies (Yes=1)	-0.06	0.117	-0.54	0.591
Education of women (Yes=1)	0.05	0.099	0.49	0.622
$1\rightarrow$ 1				
Constant	0.19	0.211	0.90	0.366
Age at marriage ($\leq 14 = 1$)	-0.02	0.133	-0.12	0.902
Wanted pregnancy (Yes=1)	0.01	0.140	0.04	0.969
Visit to HC (yes=1)	-1.40	0.135	10.37	0.000
Previous pregnancies (Yes=1)	0.13	0.155	0.85	0.396
Education of women (Yes=1)	0.40	0.132	3.00	0.003
$2\rightarrow$ 2				
Constant	-0.51	0.616	-0.83	0.405
Age at marriage ($\leq 14 = 1$)	0.10	0.365	0.27	0.790
Wanted pregnancy (Yes=1)	0.45	0.381	1.19	0.234
Visit to HC (yes=1)	-0.47	0.391	-1.19	0.233
Previous pregnancies (Yes=1)	1.15	0.421	2.73	0.006
Education of women (Yes=1)	0.32	0.374	0.85	0.396
Chi-square (p-value)	973.91 (0.0000)			

transitions of the types $0 \rightarrow 1$, $1 \rightarrow 1$ and $2 \rightarrow 2$ are considered in Table 2. It is observed from the results that visit to healthcare facilities is positively associated with $0 \rightarrow 1$, and negatively associated with $1 \rightarrow 1$ and number of previous pregnancies is positively associated with the transition of the type $2 \rightarrow 2$.

6 Conclusion

The discrete time Markov models are used in characterizing the pattern of transition in the disease states. A number of researchers have used different Markov models. This paper uses a simple method of linking the transition probabilities with their potential risk factors by employing the logistic regression model. Muenz and Rubinstein (1985) proposed a covariate dependent model for two intercommunicating states. In this paper, a three state Markov model is used to show the covariate dependence, and the proposed model can be generalized for any finite number of states. The likelihood function, estimation procedure and test procedures are discussed in the paper. The proposed method generalizes the estimation and test procedures as well as the utility of the two-state covariate dependent Markov model for any finite number of states. An application is included in this paper to illustrate the use of the proposed model for real life problems.

References

[1] Albert, P.S. (1994). A Markov Model for Sequence of Ordinal Data from a relapsing-remitting disease. *Biometrics,* **50** : 51-60.

[2] Albert, P.S. and Myron, A.W. (1998). A two State Markov Chain for Heterogeneous Transitional Data: A Quasilikelihood Approach. *Statist. Med.*, **17**: 1481-1493.

[3] Anderson, T.W and Goodman, L. (1957). Statistical Inference about Markov Chains. *Annals of Mathematical Statistics,* **28** : 89-110.

[4] Billingsley, P. (1961). Statistical Inference for Markov Processes. The University of Chicago Press: Chicago, USA.

[5] Hosmer, D.W. and Lemeshow, S. (1989). *Applied Logistic Regression.* John Willy and Sons: New York, USA.

[6] Kalbfleisch, J.D. and Lawless, J.F. (1985). The Analysis of Panel Data Under a Markov Assumption. *Journal of American Statistical Association,* **80**: 863-871.

[7] Korn, E.L. and Whittemore, A.S. (1979). Methods of Analyzing Panel Studies of Acute Health Effects of Air Pollution. *Biometrics*, **35**: 795-802.

[8] Muenz, L.R. and Rubinstein, L.V. (1985). Markov Models for Covariate Dependence of Binary Sequences. *Biometrics,* **41**: 91-101.

[9] Prentice, R. and Gloeckler, L. (1978). Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data. *Biometrics,* **34**: 57-67.

[10] Raftery, A. and Tavare, S. (1994). Estimating and Modeling Repeated Patterns in Higher Order Markov Chains with the Mixture Transition Distribution Model. *Appl. Statist.,* **43** (1) :179-199.

[11] Regier, M.H. (1968). A Two State Markov Model for Behavior Change. *Journal of American Statistical Association*; **63**: 993-999.

BLANK PAGE