



Received: 31 July 2017
Accepted: 20 December 2017
First Published: 05 January 2018

*Corresponding author: M. Ataharul Islam, ISRT, University of Dhaka, Dhaka 1000, Bangladesh
E-mail: mataharul@yahoo.com

Reviewing editor:
Yan Sun, Utah State University, USA

Additional information is available at the end of the article

STATISTICS | RESEARCH ARTICLE

Goodness of fit test for higher order binary Markov chain models

Mahboobeh Zangeneh Sirdari¹ and M. Ataharul Islam^{2*}

Abstract: When the interest is in making statements about change based on repeated measurements of discrete data, one way to do so is using Markov chain models. Goodness of fit test to find a good model is very important in analyzing the underlying patterns and relationships in the repeated measures data. To test for the various associations in the models, the likelihood ratio and Wald tests are used. However, it has been observed that the efficient score tests can provide equally good tests and can provide an easier alternative. In this paper, we provide an extension of Tsiatis method for goodness of fit test on higher order Markov chains. In our method, we follow the approach of Tsiatis goodness of fit test in logistic regression models. New method provided in this paper is applied to real-life data to examine the suitability of the techniques.

Subjects: Statistics & Probability; Probability Theory & Applications; Statistics; Mathematical Statistics

Keywords: Markov chain; goodness of fit test; efficient score tests

1. Introduction

Markov chain models are used in various applied fields, such as time series analysis, longitudinal studies, life data, environmental problems. The behavior of a Markov chain depends on the transition

ABOUT THE AUTHORS

M. Ataharul Islam is currently the QM Husain professor at the Institute of Statistical Research, University of Dhaka, Bangladesh. He is a recipient of the Pauline Stitt Award, Western North American Region (WNAR) Biometric Society Award for content and writing, East West Center Honor Award, University Grants Commission Award for book and research, and the Ibrahim Memorial Gold Medal for research. He has published more than 100 papers in international journals on various topics, mainly on longitudinal and repeated measures data including multistate and multistage hazards model, statistical modeling, Markov models with covariate dependence, generalized linear models, conditional and joint models for correlated outcomes. He supervised research of more than 100 students at MS and PhD levels. He authored books on Analysis of Repeated Measures Data, Markov models, edited one book jointly and contributed chapters in several books.

PUBLIC INTEREST STATEMENT

In this paper, an important test procedure arising from repeated measurements of discrete data have been introduced. In real-life situations, the change in the status of a disease or other outcome variables need to be examined. These changes need to be studied to understand the underlying factors influencing transitions in the status during specified time intervals. We can employ Markov models in order to find the relationships between the risk factors and outcome variables. The models can be of first or higher orders. One formidable challenge in employing these models is to confirm goodness of the model for first or higher order. This paper provides a simple test that can be used to test goodness of fit of higher order Markov models with covariate dependence for binary data. The proposed test procedure appears to be very useful technique in providing the goodness of fit of higher order binary Markov chains.

matrix, which contains transitional probabilities. In most practical studies, the transition matrix is unknown and needs to be estimated. Several methods are available for the estimation and test procedure of transition probabilities. However, most researchers have worked primarily on the estimation of parameters and only a few reports on test procedures. One of the most important tests on Markov chain models is the stationarity of transition probabilities and the goodness of fit of Markov chain models. This section presents a brief summary of the tests performed on Markov chain.

Anderson and Goodman (1957) obtained the maximum likelihood estimates and their asymptotic distribution for the transition probabilities in a Markov chain of arbitrary order with repeated observations of the chain. The likelihood ratio tests and chi-square tests used in contingency tables were obtained for testing these hypotheses. Billingsley (1961) used Whittle's formula, chi-square, and maximum likelihood methods to test for stationarity and order of the higher order Markov chain. McQueen and Thorley (1991) used Markov chain to analyze annual stock returns. Albert (1994) proposed a class of Markov models for analyzing sequences of ordinal data from a relapsing-remitting disease, where the state space was expanded to include information about the ordinal severity score as well as the relapsing-remitting status. He proposed a parameterization that can reduce the number of parameters. It is noteworthy that most of these research works have been conducted for estimating parameters based on the first-order Markov chain. Recently, several new methods for higher order Markov chains have been reported, where the estimation and test procedures became quite complex due to the increased order of the models (Chowdhury, Islam, Shah, & Al-Enezi, 2005; Islam & Chowdhury, 2006; Islam, Chowdhury, & Briollais, 2012; Rahman & Islam, 2007).

However, less effort is given towards studying the field of covariate-dependent Markov models (Muenz & Rubinstein, 1985; Yi, He, & Liang, 2009). In this paper, a test procedure for the goodness of fit of a binary Markov chain model is proposed by extending Tsiatis' procedure (Tsiatis, 1980). The proposed test was extended for the second- and higher order of the Markov chain model. The efficient score test was used for testing null hypotheses, which only required the estimate of parameters under true null hypothesis. The proposed model and test procedures were thoroughly examined using a set of data for the elderly population and employing simulations.

Sirdari, Islam, and Awang (2013) proposed the goodness of fit test for higher order binary Markov chain models based on marginal distribution. The problem with this proposal was that marginal distribution has limited assumptions because of the correlations between variables, which are not easy to estimate. Thus, the proposed model in this study was based on the conditional transition probabilities, which means that there is no correlation between variables.

2. A brief overview of the test proposed by Tsiatis (1980)

Tsiatis (1980) proposed a goodness of fit test for the logistic regression model. In terms of binary data analysis, this model relates the probability of a response to a set of covariates (χ_1, \dots, χ_p) according to Equation (2.1):

$$\log \left\{ \frac{p_x}{1-p_x} \right\} = \beta' \chi, \quad p_x = \frac{\exp(\beta' \chi)}{1 + \exp(\beta' \chi)}, \quad (2.1)$$

where p_x denotes the conditional probability of response given by the vector, $\chi = (\chi_1, \dots, \chi_p)$, $\chi_0 = 1$, and $\beta' = (\beta_0, \dots, \beta_p)$ denotes the regression coefficients. The space of covariates (χ_1, \dots, χ_p) is partitioned into k distinct region in p -dimensional space, denoted by R_1, \dots, R_k . The indicator functions, $\mathbf{I}^{(j)}$ ($j = 1, \dots, k$), are defined by $\mathbf{I}^{(j)} = 1$ if $(\chi_1, \dots, \chi_p) \in R_j$ and $\mathbf{I}^{(j)} = 0$.

Tsiatis considered the following model in Equation (2.2):

$$\log \left\{ p_x / (1 - p_x) \right\} = \beta' \chi + \gamma' \mathbf{I}, \quad (2.2)$$

where $\mathbf{I}' = (I^{(1)}, \dots, I^{(k)})$ and $\gamma' = (\gamma_1, \dots, \gamma_k)$. The goodness of fit test consists of testing the hypothesis $H_0: \gamma_1 = \dots = \gamma_k = 0$.

This test is based on the efficient score test, as represented by Equation 2.3:

$$T = \mathbf{Z}' \mathbf{V}^{-1} \mathbf{Z}, \quad (2.3)$$

where \mathbf{Z}' is the k -dimensional vector $(\partial l / \partial \gamma_1, \dots, \partial l / \partial \gamma_k)$ and l denotes the log-likelihood. The $k \times k$ matrix, \mathbf{V} is equal to:

$$\mathbf{V} = \mathbf{A} - \mathbf{BC}^{-1} \mathbf{B}',$$

where

$$A_{jj'} = -\partial^2 l / \partial \gamma_j \partial \gamma_{j'} \quad (j, j' = 1, \dots, k),$$

$$B_{jj'} = -\partial^2 l / \partial \gamma_j \partial \beta_{j'} \quad (j = 1, \dots, k; j' = 0, \dots, p),$$

$$C_{jj'} = -\partial^2 l / \partial \beta_j \partial \beta_{j'} \quad (j, j' = 0, \dots, p)$$

All previous terms were evaluated at $\gamma = 0$ and $\beta_j = \hat{\beta}_j$, where $\hat{\beta}_j$ is the maximum likelihood estimate of the parameters when is true.

3. Goodness of fit test of first-order Markov chains

Consider the case of a single stationary process, (Y_1, \dots, Y_T) , generated by a binary Markov chain that uses values of 0 and 1. The transition matrix is defined by:

$$P = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} = \begin{bmatrix} 1 - p_{01} & p_{01} \\ 1 - p_{11} & p_{11} \end{bmatrix}$$

where $p_{jt} = \Pr(Y_t = 1 | Y_{t-1} = j); j = 0, 1, t = 1, \dots, T$.

The transition probabilities, p_{jt} , can be modeled using logistic regression, as shown by Model (3.1):

$$\text{logit}(p_{jt}) = \beta_j' \chi_t, \quad p_{jt} = \frac{\exp(\beta_j' \chi_t)}{1 + \exp(\beta_j' \chi_t)}. \quad (3.1)$$

Vector χ_t contains covariates and it is equal to $\chi_t = (1, \chi_{t1}, \dots, \chi_{tp})$. β_j is the vector of parameters, $\beta_j = (\beta_{j0}, \beta_{j1}, \dots, \beta_{jp})$. The likelihood function that corresponds to Model (3.1) is:

$$L = \prod_t \prod_{j=0}^1 (1 - p_{jt})^{n_{j0t}} p_{jt}^{n_{j1t}}$$

where n_{00t} , n_{01t} , n_{10t} , and n_{11t} are the number of transitions of each type observed at time t . The log-likelihood is as shown by Equation (3.2):

$$l = \sum_{t=1}^T \sum_{j=0}^1 \left\{ n_{j1t} \beta_j' \chi_t - (n_{j0t} + n_{j1t}) \ln [1 + \exp(\beta_j' \chi_t)] \right\}. \quad (3.2)$$

The log-likelihood can be shown as, $l = \ln L = \ln L_0 + \ln L_1$, where

$$\ln L_0 = \sum_{t=1}^T \{n_{01t} \beta'_0 \chi_t - (n_{00t} + n_{01t}) \ln [1 + \exp(\beta'_0 \chi_t)]\},$$

$$\ln L_1 = \sum_{t=1}^T \{n_{11t} \beta'_1 \chi_t - (n_{10t} + n_{11t}) \ln [1 + \exp(\beta'_1 \chi_t)]\}.$$

It was assumed that the space of covariate $(\chi_{t1}, \dots, \chi_{tp})$ was partitioned into G distinct regions in p -dimensional space, denoted by R_1, \dots, R_G . The indicator functions, $I_t^{(k)}$ ($k = 1, \dots, G$), are defined by $I_t^{(k)} = 1$ if $(\chi_{t1}, \dots, \chi_{tp}) \in R_k$ and $I_t^{(k)} = 0$.

Then, for a binary Markov chain, the following Model (3.3) was considered:

$$\text{logit}(p_{jt}) = \beta'_j \chi_t + \gamma'_j \mathbf{I}_t, \tag{3.3}$$

where $\mathbf{I}_t = (I_t^{(1)}, \dots, I_t^{(G)})$ and $\gamma'_j = (\gamma_{j1}, \dots, \gamma_{jG})$ is an arbitrary covariate vector. This test was performed by testing the null hypothesis, $H_0: \gamma_{j1} = \dots = \gamma_{jG} = 0$. This hypothesis was proposed by partitioning the space of covariates into distinct regions and calculating a test statistic, which was a quadratic form of the observed counts, excluding the expected counts.

The efficient score test and the likelihood ratio test were also used. Both statistics have asymptotic chi-square distribution, with G degrees of freedom, as proven by Rao (1973). The current test used in this study was based on the efficient score test because it only requires an estimate of β_j under the null hypothesis, whereas the likelihood ratio statistics needs an estimate of γ_j under the alternative model. The test statistics is defined by Equation (3.4):

$$T = \mathbf{Z}' \mathbf{V}^{-1} \mathbf{Z}, \tag{3.4}$$

where $\mathbf{Z}' = (\mathbf{Z}'_0 \quad \mathbf{Z}'_1)$ and \mathbf{Z}'_j , $j = 0, 1$ is the G -dimensional vector $(\partial l / \partial \gamma_{j1}, \dots, \partial l / \partial \gamma_{jG})$. The matrix, \mathbf{V} , is equal to:

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_0 & 0 \\ 0 & \mathbf{V}_1 \end{pmatrix}$$

and the $G \times G$ matrix, \mathbf{V}_j , $j = 0, 1$

$$\mathbf{V}_j = \mathbf{A}_j - \mathbf{B}_j \mathbf{C}_j^{-1} \mathbf{B}'_j,$$

where

$$\mathbf{A}_{jkk'} = -\partial^2 l / \partial \gamma_{jk} \partial \gamma_{jk'} \quad (k, k' = 1, \dots, G),$$

$$\mathbf{B}_{jkk'} = -\partial^2 l / \partial \gamma_{jk} \partial \beta_{jk'} \quad (k = 1, \dots, G; k' = 0, \dots, p),$$

$$\mathbf{C}_{jkk'} = -\partial^2 l / \partial \beta_{jk} \partial \beta_{jk'} \quad (k, k' = 0, \dots, p).$$

All previous terms were evaluated at $\gamma_j = 0$ and $\beta = \hat{\beta}$, where $\hat{\beta}$ is the maximum likelihood estimate of the parameters when H_0 is true. Using the standard likelihood theory, a test could be extracted in the quadratic form of observed counts minus expected counts, and whose large sample properties are easily established.

It is evident that the log-likelihood for Model (3.3) can be achieved by inserting Equation (3.2), as follows:

$$l = \sum_{t=1}^T \sum_{j=0}^1 \left[n_{j1t} (\beta'_j \chi_t + \gamma'_j I_t) - (n_{j0t} + n_{j1t}) \ln \left\{ 1 + \exp (\beta'_j \chi_t + \gamma'_j I_t) \right\} \right] = \ln L_0 + \ln L_1 = l_0 + l_1.$$

The k th element of vector χ_t used in the computation of Equation (3.4) is the partial derivative of l_j , $j = 0, 1$, with respect to γ_{jk} at $\gamma_j = 0$ and $\beta_j = \hat{\beta}_j$,

$$\sum_{t=1}^T n_{j1t} I_t^{(k)} - \sum_{t=1}^T (n_{j0t} + n_{j1t}) I_t^{(k)} \left[\frac{\exp (\beta'_j \chi_t)}{\left\{ 1 + \exp (\beta'_j \chi_t) \right\}} \right] = O_{jk} - E_{jk},$$

where O_{jk} and E_{jk} are the observed and expected numbers of responses in the k th region. Therefore, Equation (3.4) is the quadratic form of the vector of observed counts minus expected counts.

Quantities necessary for computing the covariance matrix, \mathbf{V}_j , $j = 0, 1$ are as follows:

$$A_{jkk'} = \begin{cases} \sum_{\xi_k} (n_{j0t} + n_{j1t}) \hat{p}_{jt} (1 - \hat{p}_{jt}) & k = k' \\ 0 & k \neq k'; k, k' = 1, \dots, G \end{cases}$$

$$B_{jkk'} = \sum_{\xi_k} (n_{j0t} + n_{j1t}) \chi_{tk'} \hat{p}_{jt} (1 - \hat{p}_{jt}) \quad (k = 1, \dots, G; k' = 0, \dots, p),$$

$$C_{jkk'} = \sum_{t=1}^T (n_{j0t} + n_{j1t}) \chi_{tk} \chi_{tk'} \hat{p}_{jt} (1 - \hat{p}_{jt}) \quad (k, k' = 0, \dots, p),$$

where, ξ_k denotes the set of indices t , such that

$$(\chi_{t1}, \dots, \chi_{tp}) \in R_k, \quad \hat{p}_{jt} = \exp (\hat{\beta}'_j \chi_t) / \left\{ 1 + \exp (\hat{\beta}'_j \chi_t) \right\}.$$

4. Extension of the model for higher-order Markov chains

Consider the n th-order Markov model for times, $t - n$, $t - (n - 1)$, \dots , $t - 1$, and t , with transition matrix, \mathbf{P} , and its components:

$$p_{r \dots sjt} = \Pr (Y_t = 1 | Y_{t-n} = r, \dots, Y_{t-2} = s, Y_{t-1} = j); j, s, r = 0, 1, t = 1, \dots, T.$$

The logistic regression model for $p_{r \dots sjt}$ is:

$$\text{logit} (p_{r \dots sjt}) = \beta'_{r \dots sj} \chi_t, \quad p_{r \dots sjt} = \frac{\exp (\beta'_{r \dots sj} \chi_t)}{1 + \exp (\beta'_{r \dots sj} \chi_t)},$$

where vector, $\chi_t = (1, x_{t1}, \dots, x_{tp})$, contains covariates and $\beta_{r...sj} = (\beta_{r...sj0}, \beta_{r...sj1}, \dots, \beta_{r...sjp})$ is the vector of parameters.

The likelihood function corresponding to this model is as follows:

$$L = \prod_t \prod_{j=0}^1 \prod_{s=0}^1 \dots \prod_{r=0}^1 (1 - p_{r...sjt})^{n_{r...sj0t}} p_{r...sjt}^{n_{r...sj1t}}$$

The log-likelihood is:

$$l = \sum_{t=1}^T \sum_{j=0}^1 \sum_{s=0}^1 \dots \sum_{r=0}^1 \left\{ n_{r...sj1t} \beta'_{r...sj} \chi_t - (n_{r...sj0t} + n_{r...sj1t}) \ln [1 + \exp(\beta'_{r...sj} \chi_t)] \right\} = \sum_{r,s,j=0}^1 \ln L_{r...sj},$$

where for $r, s, j = 0, 1$,

$$\ln L_{r...sj} = \sum_{t=1}^T \left\{ n_{r...sj1t} \beta'_{r...sj} \chi_t - (n_{r...sj0t} + n_{r...sj1t}) \ln [1 + \exp(\beta'_{r...sj} \chi_t)] \right\}.$$

Then, Model (3.3) can be extended for order n , which can be written as shown by Equation (4.1):

$$\text{logit}(p_{r...sjt}) = \beta'_{r...sj} \chi_t + \gamma'_{r...sj} \mathbf{I}_t; \quad r, s, j = 0, 1. \tag{4.1}$$

The related null hypothesis is $H_0: \gamma_{r...sj1} = \dots = \gamma_{r...sjG} = 0$, and the test statistic is shown by Equation (4.2):

$$T = \mathbf{Z}' \mathbf{V}^{-1} \mathbf{Z}, \tag{4.2}$$

where \mathbf{Z} is a 2^n -dimensional vector with elements of

$$\mathbf{Z}'_{r...sj} = \left(\partial l / \partial \gamma_{r...sj1}, \dots, \partial l / \partial \gamma_{r...sjG} \right), \quad r, s, j = 0, 1.$$

The matrix, \mathbf{V} is the $2^n \times 2^n$ diagonal matrix, with components of $G \times G$ matrix. $\mathbf{V}_{r...sj}; r, s, j = 0, 1$;

$$\mathbf{V}_{r...sj} = \mathbf{A}_{r...sj} - \mathbf{B}_{r...sj} \mathbf{C}_{r...sj}^{-1} \mathbf{B}'_{r...sj},$$

where

$$\mathbf{A}_{r...sjkk'} = -\partial^2 l / \partial \gamma_{r...sjk} \partial \gamma_{r...sjk'} \quad (k, k' = 1, \dots, G),$$

$$\mathbf{B}_{r...sjkk'} = -\partial^2 l / \partial \gamma_{r...sjk} \partial \beta_{r...sjk'} \quad (k = 1, \dots, G; k' = 0, \dots, p),$$

$$\mathbf{C}_{r...sjkk'} = -\partial^2 l / \partial \beta_{r...sjk} \partial \beta_{r...sjk'} \quad (k, k' = 0, \dots, p), \quad (r, s, j = 0, 1).$$

The log-likelihood based on Model (4.1) is as follows:

$$l = \sum_{t=1}^T \sum_{j=0}^1 \sum_{s=0}^1 \cdots \sum_{r=0}^1 \left\{ n_{r\dots sj1t} \left(\beta'_{r\dots sj} \chi_t + \gamma'_{r\dots sj} \mathbf{I}_t \right) - \left(n_{r\dots sj0t} + n_{r\dots sj1t} \right) \ln \left[1 + \exp \left(\beta'_{r\dots sj} \chi_t + \gamma'_{r\dots sj} \mathbf{I}_t \right) \right] \right\}$$

$$= \sum_{r,s,j=0}^1 \ln L_{r\dots sj} = \sum_{r,s,j=0}^1 l_{r\dots sj}$$

The partial derivative of $l_{r\dots sj}$, $l, s, j = 0, 1$ with respect to $\gamma_{r\dots sjk}$, $r, s, j = 0, 1$ at $\gamma_{r\dots sj} = 0$ and $\beta_{r\dots sj} = \hat{\beta}_{r\dots sj}$

$$\sum_{i=1}^n n_{r\dots sj1t} I_t^{(k)} - \sum_{i=1}^n \left(n_{r\dots sj0t} + n_{r\dots sj1t} \right) I_t^{(k)} \left[\frac{\exp \left(\beta'_{r\dots sj} \chi_t \right)}{\left\{ 1 + \exp \left(\beta'_{r\dots sj} \chi_t \right) \right\}} \right] = O_{r\dots sjk} - E_{r\dots sjk}$$

5. Application

We applied the proposed test on the Health and Retirement Study (HRS) (2009) data to demonstrate its application. This is a longitudinal household survey data-set for the study of retirement and health among the elderly in the United States. The RAND Centre collected these data to study aging, with funding and support from the National Institute on Aging (NIA) and the Social Security Administration (SSA). These data were collected from 1992 to 2006 in eight waves for 30,405 people. We considered individuals who attended the program in 1992 and then, followed up until 2006. The study was about depression among individuals (0 for no depression and 1 for depression), and age (yearly), gender (0 for male and 1 for female), body mass index (BMI), and drinking (0 for not drinking and 1 for drinking), which were considered as covariates. The space of covariate $(\chi_{t1}, \dots, \chi_{tp})$ was partitioned into four distinct regions: (male and not drinking); (male and drinking); (female and not drinking); and (female and drinking). Some of these variables may contain missing values because the referenced person did not respond to the waves. Thus, we had to drop the ID of individuals from all waves if there were missing values for these covariates. There were 668 missing values in the covariates, which included 353 IDs, i.e. these individuals responded for the outcome variable, but not for the covariates. Thus, 353 IDs were dropped from the data in this work. Additionally, S-Plus functions modified by Chowdhury et al. (2005) were developed and used to estimate the parameters of the model. The Newton-Raphson method was used in this program for parameter estimation.

Table 1 shows the different types of transition counts for the first- and second-order transitions. Meanwhile, Table 2 shows the estimated values for the covariate-dependent Markov models for different types of transitions. The results are for the first- and second-order Markov models.

Billingsley's chi-square statistics were computed using $\sum_{ij} \left(f_{ij} - f_i p_{ij} \right)^2 / \left(f_i p_{ij} \right)$, and Tsiatis' statistics were estimated using Equations (3.4) and (4.2). The results showed that the data satisfied the models for the first- and second-order Markov chains. Both Billingsley's and Tsiatis' statistics showed

Table 1. Transition counts of Markov chain of depression data for the first- and second-order

First-order	Transition time		t	
		t - 1	0	1
		0	3,951	1,455
		1	1,930	257
Second-order	t - 2	t - 1	t	
	0	0	3,473	1,396
	1	0	680	221
	0	1	269	29
	1	1	568	151

Table 2. Estimates of parameters of covariate-dependent first- and second-order Markov models and testing the goodness of fit

First-order				
<i>Transition type</i>	<i>Covariates</i>	<i>Estimated value</i>	<i>s.e.</i>	<i>p-value</i>
0→1	Constant	4.631	0.349	0.000
	Age	-0.096	0.005	0.000
	Sex	0.129	0.066	0.051
	BMI	0.011	0.006	0.078
	Drinking	-0.215	0.066	0.0011
1→1	Constant	8.378	0.825	0.000
	Age	-0.118	0.012	0.000
	Sex	0.019	0.148	0.896
	BMI	0.007	0.012	0.577
	Drinking	0.552	0.148	0.0002
Billingsley's chi-square		3.94E-13	(p-value = 0.999)	
Proposed test statistics		1.207	(p-value = 0.997)	
Second-order				
<i>Transition type</i>	<i>Covariate</i>	<i>Estimated value</i>	<i>s.e.</i>	<i>p-value</i>
0→0→1	Constant	4.360	0.385	0.000
	Age	-0.090	0.005	0.000
	Sex	0.224	0.068	0.0011
	BMI	0.009	0.006	0.153
	Drinking	-0.282	0.067	0.0003
1→0→1	Constant	-2.996	0.999	0.003
	Age	0.020	0.015	0.178
	Sex	0.335	0.174	0.053
	BMI	0.026	0.016	0.104
	Drinking	-0.278	0.160	0.082
0→1→1	Constant	2.261	2.250	0.315
	Age	0.006	0.034	0.870
	Sex	0.195	0.423	0.646
	BMI	-0.015	0.036	0.682
	Drinking	-0.101	0.410	0.805
1→1→1	Constant	9.295	1.217	0.000
	Age	-0.142	0.019	0.000
	Sex	0.026	0.220	0.907
	BMI	0.007	0.015	0.670
	Drinking	0.513	0.222	0.021
Billingsley's chi-square		2.13E-08	(p-value = 0.999)	
Proposed test statistics		2.057	(p-value = 0.999)	

similar results. However, Billingsley's test statistics does not depend on covariates. Thus, it was used in this study to compare the results with results of the extended test based on Tsatis' statistics.

Estimates of the parameters for the first-order transitions demonstrated negative association between the transitions from no depression to depression with age and drinking, while positive associations were obtained with BMI and sex (females have higher risks). Those who did not change their

status from depression were found to be associated negatively with age and positively with drinking. Both the Billingsley and the proposed tests showed that the first-order models can be accepted. Similarly, the second-order models indicated that age and drinking were negatively associated, while sex was positively associated for the 0-0-1 type of transition. Sex and BMI were positively associated and drinking was negatively associated for the 1-0-1 transition type, while age was negatively associated, but drinking was positively associated for the 1-1-1 transition type. Interestingly, the second-order models also appeared to have good fit in favor of the null hypothesis. These results demonstrated that both the first- and second-order models could be employed for the given set of data.

6. Simulation

Data generated by the techniques provided by Ghosh and Mukerjee, and Leisch et al. was used to examine the suitability of the proposed models. In these techniques, `bindata` package in R were employed for generating correlated binary data. First, data were generated from the multivariate normal random variables, and then, they were transformed into binary data. In this study, two variables were generated as the outcome variables at time t and $t - 1$ for the first-order Markov model, with various combinations of probabilities of occurring 1 and 0 to obtain different correlations. These results were used to compare the models under independent and selected values of measure of association. For models 1, 2, and 3, the data were generated based on correlation of 0.4 between the outcome variables at time $t - 1$ and t for the first-order, and at time $t - 2$, $t - 1$, and t for the second-order. Similarly, models 4, 5, and 6 were generated with correlation of 0, while models 7, 8, and 9 considered correlation of -0.4 . For each model, four covariates were also generated, corresponding to the correlated response variables by considering different correlations with the outcome variables. These estimates and tests were repeated 500 times for all models, and for sample sizes of 250, 500, and 1,000 for different correlations between outcome variables. The models that were used for this simulation study were different applications of the conditional model verified in Model (3.3). The extended Tsiatis' test for the first-order Markov model, as shown by Equation (4.2), had involved covariate patterns. Nonetheless, Billingsley's test, which was represented by $\sum_{ij} (f_{ij} - f_i p_{ij})^2 / (f_i p_{ij})$, was used to compare the obtained results, with and without covariates. In other words, the Markov models were estimated using covariates and were employed in this test.

Table 3 shows the simulation results for the first-order model, which included frequencies by transition type, correlation between outcome variables in the bivariate Bernoulli population, average estimates of the parameters, and the number of rejected hypotheses in 500 times of simulation for these models using Billingsley's test and the proposed test as an extension of Tsiatis' test. Acceptance of the null hypothesis, $H_0: \gamma_{j1} = \dots = \gamma_{jG} = 0$, would indicate a good fit of Model (4.1) to the data. The percentage of rejection for Billingsley's test was 0 because the test procedure did not consider any covariate. However, in the proposed extension, models with covariate dependence were used. Hence, it can be concluded that the proposed test statistics depended on the covariate-dependant transition probabilities, where the selection of appropriate variables in the model may influence the goodness of fit, to a large extent. This observation implied that the proposed test may display deviations for a good fit in some instances. In other words, the goodness of fit test proposed by this study, as an extension of the Tsiatis' test, depended on the model's specifications in terms of the explanatory power of the selected variables. Based on the estimated covariates for the first-order transition from 0 to 1, we observed a positive association with variable 1, and a transition from 1 to 1, which has negative association with variable 1 for all models, except with model 4 (sample size of 250, with correlation of 0). The rejection percentage had varied for the first-order model, mainly in the range of 4.-6.6%. This result showed that the proposed test method was satisfactory for different sizes of samples, with different correlation of outcome variables based on the first-order Markov chain model.

The simulation results for the second-order model are given in Table 4. The table shows the number of transition types, correlation between outcome variables, average estimates of the parameters, and the number of rejected hypotheses, $H_0: \gamma_{r\dots sj1} = \dots = \gamma_{r\dots sjG} = 0$, in 500 times of simulations

Transition type	Model 1—size 250		Model 2—size 500		Model 3—size 1000		Model 4—size 250		Model 5—size 500	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
00	61	0.115	121	0.025	243	0.002	50	0.130	100	0.040
01	65	0.011	129	0.000	257	0.000	50	0.143	100	0.043
10	14	0.438	29	0.383	58	0.304	75	0.396	150	0.313
11	110	0.482	221	0.469	442	0.464	75	0.321	150	0.237
Correlation of response variables	0.4	0.423	0.4	0.380	0.4	0.293	0	0.263	0	0.126
Estimates of Parameters	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
0 to 1	Con-stant	-0.868	-0.843	0.025	-0.830	0.002	-0.916	0.130	-0.877	0.040
	V1	1.422	1.386	0.000	1.360	0.000	0.972	0.143	0.942	0.043
	V2	0.301	0.290	0.383	0.287	0.304	0.341	0.396	0.357	0.313
	V3	0.093	0.109	0.469	0.101	0.464	0.517	0.321	0.469	0.237
	V4	0.305	0.262	0.380	0.259	0.293	0.759	0.263	0.751	0.126
1 to 1	Con-stant	-1.181	-1.114	0.084	-1.044	0.019	1.233	0.045	1.200	0.005
	V1	-1.776	-1.696	0.005	-1.685	0.000	-0.914	0.072	-0.907	0.012
	V2	-0.365	-0.257	0.453	-0.250	0.418	-0.364	0.352	-0.350	0.273
	V3	-0.013	-0.037	0.483	-0.061	0.468	-0.496	0.292	-0.491	0.161
	V4	-0.252	-0.253	0.448	-0.285	0.376	-0.693	0.157	-0.665	0.058
Billingsley	0.008	0.951	0.010	0.938	0.019	0.907	2.50E-05	0.999	1.06E-05	0.999
No. of tests accepting H_0	500	500	500	500	500	500	500	500	500	500
Proposed test	8.646	0.462	8.267	0.481	8.172	0.481	8.159	0.486	8.403	0.470
No. of tests accepting H_0	460	467	476	476	476	476	474	474	470	470
Proportion of rejection of H_0	40/500	33/500	33/500	33/500	24/500	24/500	26/500	26/500	30/500	30/500
Transition										
00			201		26		51		102	
01			200		99		200		399	
10			300		75		148		298	
11			300		50		101		201	
Correlation of response variables			0		-0.4		-0.4		-0.4	
Estimates of Parameters	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
0 to 1	Constant	-0.875	-0.875	0.003	0.314	0.028	0.489	0.202	0.492	0.081
	V1	0.926	0.926	0.004	1.797	0.028	1.707	0.002	1.688	0.000
	V2	0.351	0.351	0.215	2.183	0.036	1.961	0.002	1.931	0.000
	V3	0.470	0.470	0.112	0.709	0.274	0.666	0.172	0.635	0.066
	V4	0.743	0.743	0.028	-0.715	0.269	-0.666	0.166	-0.653	0.063
1 to 1	Constant	1.190	1.190	0.000	2.463	0.003	2.315	0.000	2.291	0.000
	V1	-0.885	-0.885	0.000	-1.731	0.013	-1.635	0.001	-1.613	0.000
	V2	-0.342	-0.342	0.154	-1.621	0.011	-1.546	0.000	-1.529	0.000
	V3	-0.491	-0.491	0.060	-0.645	0.242	-0.605	0.153	-0.611	0.042
	V4	-0.664	-0.664	0.007	0.666	0.247	0.642	0.116	0.657	0.024
Billingsley	4.79E-06	0.999	0.011	0.935	0.011	0.935	0.011	0.934	0.011	0.923
No. of tests accepting H_0	500	500	500	500	500	500	500	500	500	500
Proposed test	8.167	0.489	8.167	0.467	8.383	0.467	8.439	0.468	8.448	0.463
No. of tests accepting H_0	472	472	471	471	471	471	475	475	473	473
Proportion of rejection of H_0	28/500	28/500	29/500	29/500	29/500	29/500	25/500	25/500	27/500	27/500

Table 4. Five hundred simulations for obtaining the estimates of associations based on the proposed second-order models

Transition	Model 1—size 250			Model 2—size 500			Model 3—size 1000			
	Estimate	p-value		Estimate	p-value		Estimate	p-value		
0→0→1	Constant	-0.816	0.244	-0.724	0.136		-0.701	0.036		
	V1	0.670	0.313	0.585	0.258		0.559	0.133		
	V2	0.938	0.273	0.835	0.168		0.831	0.059		
	V3	0.449	0.386	0.401	0.347		0.397	0.247		
	V4	0.347	0.446	0.312	0.435		0.337	0.339		
1→0→1	Constant	-1.025	0.232	-0.989	0.101		-0.951	0.031		
	V1	0.605	0.334	0.601	0.242		0.582	0.116		
	V2	0.861	0.272	0.855	0.117		0.796	0.036		
	V3	0.425	0.404	0.397	0.369		0.419	0.235		
	V4	0.335	0.438	0.295	0.411		0.268	0.353		
0→1→1	Constant	1.035	0.235	0.990	0.101		0.938	0.031		
	V1	-0.610	0.353	-0.595	0.240		-0.566	0.132		
	V2	-0.902	0.262	-0.840	0.151		-0.806	0.049		
	V3	-0.467	0.413	-0.432	0.341		-0.391	0.272		
	V4	-0.348	0.446	-0.302	0.417		-0.304	0.364		
1→1→1	Constant	1.333	0.244	1.255	0.116		1.235	0.027		
	V1	-0.621	0.374	-0.613	0.255		-0.601	0.131		
	V2	-0.814	0.263	-0.813	0.129		-0.781	0.032		
	V3	-0.483	0.398	-0.418	0.356		-0.421	0.269		
	V4	-0.341	0.437	-0.291	0.417		-0.293	0.353		
Billingsley		2.86E-04	1.000	5.01E-05	1.000		1.66E-05	1.000		
No. of tests accepting H_0		500		500			500			
Proposed test		16.913	0.445	16.650	0.459		15.804	0.503		
No. of tests accepting H_0		475		472			482			
Proportion of rejection of H_0		25/500		28/500			18/500			
		Model 4—size 250			Model 5—size 500			Model 6—size 1000		
Transition										
000		31		63		126				
001		31		62		125				
100		32		63		124				
101		31		62		125				
010		31		63		125				

(Continued)

Table 4. (Continued)

011		31	62	125
110		31	63	125
111		32	62	125
Correlation of response variables				
<i>Estimates of parameters</i>				
0→0→1	Constant	Estimate	p-value	
	V1	-0.771	0.266	-0.704
	V2	0.615	0.343	0.144
	V3	0.912	0.288	0.260
	V4	0.412	0.400	0.163
	Constant	0.381	0.441	0.362
1→0→1	V1	-1.083	0.229	0.422
	V2	0.644	0.353	0.111
	V3	0.879	0.259	0.266
	V4	0.474	0.409	0.144
	Constant	0.326	0.448	0.327
0→1→1	V1	1.013	0.239	0.389
	V2	-0.580	0.359	0.112
	V3	-0.942	0.251	0.222
	V4	-0.434	0.409	0.177
	Constant	-0.310	0.454	0.339
1→1→1	V1	1.291	0.254	0.419
	V2	-0.611	0.377	0.101
	V3	-0.849	0.253	0.264
	V4	-0.450	0.434	0.130
	Constant	-0.330	0.437	0.360
Billingsley		3.38E-04	1.000	0.421
No. of tests accepting H_0		500	500	1.000
Proposed test		17.477	16.562	2.14E-05
No. of tests accepting H_0		559	476	500
Proportion of rejection of H_0		41/500	24/500	473
Transition		Model 7—size 250	Model 8—size 500	Model 9—size 1000
000		16	31	62
001		59	119	237
100		28	56	112
101		22	44	88
010		22	44	88
011		28	56	112
110		59	118	238
111		16	32	63
Correlation of response variables				
<i>Estimates of parameters</i>				
	Estimate	p-value	Estimate	p-value
	-0.4	-0.4	-0.4	-0.4

(Continued)

Table 4. (Continued)

0→0→1	Constant	-0.068	0.479	-0.060	0.498	-0.063	0.468
	V1	-0.051	0.487	-0.053	0.485	-0.066	0.454
	V2	-0.053	0.466	-0.072	0.500	-0.061	0.472
	V3	-0.024	0.500	-0.046	0.479	-0.012	0.501
1→0→1	V4	-0.091	0.498	-0.085	0.474	-0.092	0.479
	Constant	-0.164	0.479	-0.173	0.456	-0.179	0.367
	V1	-0.090	0.473	-0.045	0.505	-0.042	0.489
	V2	-0.078	0.503	-0.061	0.494	-0.060	0.502
0→1→1	V3	-0.056	0.493	0.034	0.514	-0.011	0.494
	V4	-0.096	0.477	-0.102	0.488	-0.085	0.481
	Constant	0.041	0.489	0.082	0.492	0.092	0.501
	V1	0.090	0.455	0.032	0.489	0.052	0.495
1→1→1	V2	0.139	0.489	0.096	0.476	0.053	0.511
	V3	0.351	0.409	0.152	0.465	0.014	0.491
	V4	0.418	0.459	0.219	0.489	0.133	0.496
	Constant	0.326	0.484	0.183	0.504	0.185	0.460
Billingsley	V1	-0.004	0.403	0.346	0.449	0.179	0.500
	V2	0.133	0.456	0.027	0.474	0.059	0.485
	V3	-0.005	0.240	0.114	0.361	0.249	0.450
	V4	0.289	0.361	0.212	0.452	0.272	0.477
No. of tests accepting H_0	9.51E-03	1.000	3.86E-03	1.000	1.12E-03	1.000	500
Proposed test	15.810	0.508	17.039	0.442	16.388	0.478	473
No. of tests accepting H_0	479	469	31/500	27/500			
Proportion of rejection of H_0	21/500						

for the models. Results for the second-order models showed that there was no association between covariates and outcome variables, which was expected because of the higher order of the underlying Markov chain. Only two models, 3 and 6, with sample size of 1000 and correlations of 0 and 0.4, have negative associations with variable 2 in different types of transitions. The range of rejection percentage of the null hypothesis for the extended Tsatis' test was 3.6–6.2% for models 1–9 in the second-order. Thus, these models were acceptable for different sizes of samples and different correlations between outcome variables. Results from Billingsley's test were compared with results from the proposed extension of Tsatis' test; the number of rejected null hypothesis was zero for Billingsley's test because it does not depend on covariates. The number of rejected null hypothesis for model 3 was the lowest.

7. Conclusion

An extension of Tsatis' test procedure was proposed in this study for first- and higher order binary Markov models by considering repeated measures. Most of the test procedures for stationarity and order of Markov chains were based on the likelihood ratio test and the usual chi-square test. We have shown a goodness of fit for the Markov chain by considering the efficient score test, which only requires estimated parameters under the null hypothesis. The utility of the proposed test has been examined, with an example for real-life data. The results indicated the suitability of these techniques. Additionally, simulation results demonstrated a Type-I error for the proposed test. In addition, the proposed test procedure was extended for higher order models and can be extended to test the order of binary Markov chains.

Funding

The authors are grateful to the HEQEP in project 3293, from the Department of Applied Statistics, East West University, and for the sponsorships by the UGC, Bangladesh and the World Bank.

Acknowledgments

We are thankful to Dr Rafiqul Islam Chowdhury for giving us permission to use the "kernopt markov.gen" program for parameter estimates. We would also like to thank the Health and Retirement Study (HRS) center for giving us permission to use RAND data in the application of the model.

Author details

Mahboobeh Zangeneh Sirdari¹

E-mail: mahboobeh@utar.edu.my

ORCID ID: <http://orcid.org/0000-0002-4733-1814>

M. Ataharul Islam²

E-mail: mataharul@yahoo.com

ORCID ID: <http://orcid.org/0000-0002-9215-1812>

¹ Department of Mathematics and Actuarial Sciences, Universiti Tunku Abdul Rahman, Sungai Long, Malaysia.

² ISRT, University of Dhaka, Dhaka 1000, Bangladesh.

Citation information

Cite this article as: Goodness of fit test for higher order binary Markov chain models, Mahboobeh Zangeneh Sirdari & M. Ataharul Islam, *Cogent Mathematics & Statistics* (2018), 5: 1421003.

References

- Albert, P. S. (1994). A Markov model for sequences of ordinal data from a relapsing-remitting disease. *Biometrics*, 50, 51–60. <https://doi.org/10.2307/2533196>
- Anderson, T. W., & Goodman, L. A. (1957). Statistical inference about Markov chains. *The Annals of Mathematical Statistics*, 28, 89–110. <https://doi.org/10.1214/aoms/1177707039>
- Billingsley, P. (1961). Statistical methods in Markov chains. *The Annals of Mathematical Statistics*, 32, 12–40. <https://doi.org/10.1214/aoms/1177705136>

- Chowdhury, R. I., Islam, M. A., Shah, M. A., & Al-Enezi, N. (2005). A computer program to estimate the parameters of covariate dependent higher order Markov model. *Computer Methods and Programs in Biomedicine*, 77, 175–181. <https://doi.org/10.1016/j.cmpb.2004.10.003>
- Health and Retirement Study (HRS). (2009). *Produced and distributed by the University of Michigan with funding from the National Institute on Aging* (Grant No. NIA U01AG09740). Waves [1–8], Year [1992–2006] [Online], [Accessed 2009]. Retrieved from World Wide Web: <http://hrsonline.isr.umich.edu/data/index.html>
- Islam, M. A., & Chowdhury, R. I. (2006). A higher order Markov model for analyzing covariate dependence. *Applied Mathematical*, 30, 477–488.
- Islam, M. A., Chowdhury, R. I., & Briollais, L. (2012). A bivariate binary model for testing dependence in outcomes. *Bulletin of the Malaysian Mathematical Sciences Society*, 35(4), 845–858.
- Mcqueen, G., & Thorley, S. (1991). Are stock returns predictable? A test using Markov chains *The Journal of Finance*, 46, 239–263. <https://doi.org/10.1111/j.1540-6261.1991.tb03751.x>
- Muenz, L. R., & Rubinstein, L. V. (1985). Markov models for covariate dependence of binary sequences. *Biometrics*, 41, 91–101. <https://doi.org/10.2307/2530646>
- Rahman Shafiqur, M., & Islam, M. A. (2007). Markov structure based logistic regression for repeated measures: An application to diabetes mellitus data. *Statistical Methodology*, 4, 448–460. <https://doi.org/10.1016/j.stamet.2007.01.006>
- Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: Wiley. <https://doi.org/10.1002/SERIES1345>
- Sirdari, M. Z., Islam, M. A., & Awang, N. (2013). A stationarity test on Markov chain models based on marginal distribution. *Statistical Methodology*, 11, 68–76. <https://doi.org/10.1016/j.stamet.2012.10.001>
- Tsatis, A. A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika*, 67, 250–251.
- Yi, G. Y., He, W., & Liang, H. (2009). Analysis of correlated binary data under partially linear single-index logistic models. *Journal of Multivariate Analysis*, 100, 278–290. <https://doi.org/10.1016/j.jmva.2008.04.012>



© 2018 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made.

You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions

You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.



***Cogent Mathematics & Statistics* (ISSN: 2574-2558) is published by Cogent OA, part of Taylor & Francis Group.**

Publishing with Cogent OA ensures:

- Immediate, universal access to your article on publication
- High visibility and discoverability via the Cogent OA website as well as Taylor & Francis Online
- Download and citation statistics for your article
- Rapid online publication
- Input from, and dialog with, expert editors and editorial boards
- Retention of full copyright of your article
- Guaranteed legacy preservation of your article
- Discounts and waivers for authors in developing regions

Submit your manuscript to a Cogent OA journal at www.CogentOA.com

