Goodness of fit test for higher-order Binary Markov chain models

Mahboobeh Zangeneh Sirdari, M. Ataharul Islam

Publisher: Cogent OA

Journal: *Cogent Mathematics*

DOI: http://doi.org/10.1080/23311835.2017.1421003

Goodness of fit test for higher-order Binary Markov chain models

Mahboobeh Zangeneh Sirdari, M. Ataharul Islam

[1]Mahboobeh Zangeneh Sirdari, [2] M. Ataharul Islam

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

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

E-mail: [1]mahboobeh@utar.edu.my, [2]mataharul@yahoo.com

Abstract

When the interest is in making statements about change based on repeated measurements of discrete data, one way to do so is by using Markov chain models. Goodness of fit is an important test for finding a good model that can be used for analyzing any underlying patterns and relationships in repeatedly measured data. To test for various associations in the models, the Likelihood Ratio and Wald tests were used. However, efficient score tests can also provide equally good tests and are easier alternatives. In this paper, we proposed an extension of Tsiatis' method for goodness of fit test on higher-order binary Markov chains. We applied Tsiatis' goodness of fit test in logistic regression models. The new method proposed in this paper was applied to real life data to examine the suitability of this technique.

1 Introduction

Markov chain models are used in various applied fields, such as time series analysis, longitudinal studies, life data, and environmental problems. The behavior of a Markov chain depends on the transition 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 (Islam et al., 2006, 2008, 2012; Chowdhury et al., 2005; Rahman and Islam, 2007).

However, less effort is given towards studying the field of covariate-dependent Markov models (Muenz and Rubinstein, 1985; Yi et al., 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 et al. (2012) 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 $(\chi_1, \ldots, \chi_p)$ according to Equation (2.1):

$$\log\{p_{\chi}/(1 - p_{\chi})\} = \boldsymbol{\beta}'\boldsymbol{\chi}, \quad p_{\boldsymbol{Z}} = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{\chi})}{\{1 + \exp(\boldsymbol{\beta}'\boldsymbol{\chi})\}}, \tag{2.1}$$

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

Tsiatis considered the following model in Equation (2.2):

$$\log\{p_{\chi}/(1 - p_{\chi})\} = \boldsymbol{\beta}'\boldsymbol{\chi} + \boldsymbol{\gamma}'\boldsymbol{I}, \tag{2.2}$$

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

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

$$T = \boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z}, \tag{2.3}$$

where, $\boldsymbol{Z}'$ is the k-dimensional vector $\left(\partial l/\partial\gamma_1, \ldots, \partial l/\partial\gamma_k\right)$ and l denotes the log-likelihood.

The $k \times k$ matrix, $\boldsymbol{V}$ is equal to:

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

where,

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

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

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

All previous terms were evaluated at $\boldsymbol{\gamma} = 0$ and $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of the parameters when H_0 is true.

3 Goodness of fit test of first-order Markov chains

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

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

where $p_{jt} = \Pr\left(Y_t = 1 \mid Y_{t-1} = j \right); j = 0, 1\ , t = 1, \ldots, T$.

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

$$\text{logit}\left(p_{jt}\right) = \boldsymbol{\beta}_j^{'}\boldsymbol{\chi}_t\ ,\ \ p_{jt} = \frac{\exp\left(\boldsymbol{\beta}_j^{'}\boldsymbol{\chi}_t\right)}{1 + \exp\left(\boldsymbol{\beta}_j^{'}\boldsymbol{\chi}_t\right)}. \tag{3.1}$$

Vector $\boldsymbol{\chi}_t$ contains covariates and it is equal to $\boldsymbol{\chi}_t = \left(1, \chi_{t1}, \ldots, \chi_{tp}\right)$. $\boldsymbol{\beta}_j$ is the vector of parameters, $\boldsymbol{\beta}_j = \left(\beta_{j0}, \beta_{j1}, \ldots, \beta_{jp}\right)$. The likelihood function that corresponds to Model (3.1) is:

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

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):

$$l = \sum_{t=1}^{T} \sum_{j=0}^{1} \left\{ n_{j1t}\boldsymbol{\beta}_j^{'}\boldsymbol{\chi}_t - \left(n_{j0t} + n_{j1t}\right) \ln\left[1 + \exp\left(\boldsymbol{\beta}_j^{'}\boldsymbol{\chi}_t\right)\right] \right\}. \tag{3.2}$$

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

$$\ln L_0 = \sum_{t=1}^{\mathrm{T}} \left\{ n_{01\mathrm{t}}\boldsymbol{\beta}_0^{'}\boldsymbol{\chi}_{\mathrm{t}} - \left(n_{00\mathrm{t}} + n_{01\mathrm{t}}\right) \ln\left[1 + \exp\left(\boldsymbol{\beta}_0^{'}\boldsymbol{\chi}_{\mathrm{t}}\right)\right] \right\},$$

$$\ln L_1 = \sum_{\mathrm{t}=1}^{\mathrm{T}} \left\{ n_{11\mathrm{t}}\boldsymbol{\beta}_1^{'}\boldsymbol{\chi}_{\mathrm{t}} - \left(n_{10\mathrm{t}} + n_{11\mathrm{t}}\right) \ln\left[1 + \exp\left(\boldsymbol{\beta}_1^{'}\boldsymbol{\chi}_{\mathrm{t}}\right)\right] \right\}.$$

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

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

$$\text{logit}\left(p_{jt}\right) = \boldsymbol{\beta}_j^{'}\boldsymbol{\chi}_t + \boldsymbol{\gamma}_j^{'}\boldsymbol{I}_t, \tag{3.3}$$

where $\boldsymbol{I}_t^{'} = \left(I_t^{(1)}, ..., I_t^{(G)}\right)$ and $\boldsymbol{\gamma}_j^{'} = (\gamma_{j1}, ..., \gamma_{jG})$ is an arbitrary covariate vector. This test was performed by testing the null hypothesis, $H_0: \gamma_{j1} = ... = \gamma_{jG} = 0$. 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 $\boldsymbol{\beta}_j$ under the null hypothesis, whereas the likelihood ratio statistics needs an estimate of $\boldsymbol{\gamma}_j$ under the alternative model. The test statistics is defined by Equation (3.4):

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

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

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

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

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

where,

$$A_{jkk'} = -\partial^2 l / \partial\gamma_{jk}\,\partial\gamma_{jk'} \quad (k, k' = 1, \ldots, G),$$

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

$$C_{jkk'} = -\partial^2 l / \partial\beta_{jk}\,\partial\beta_{jk'} \quad (k, k' = 0, \ldots, p).$$

All previous terms were evaluated at $\boldsymbol{\gamma}_j = 0$ and $\boldsymbol{\beta}_j = \hat{\boldsymbol{\beta}}_j$, where $\hat{\boldsymbol{\beta}}_j$ is the maximum likelihood estimate of the parameters when H_0 is true. By 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:

$$l = \sum_{t=1}^{T} \sum_{j=0}^{1} \left[n_{j1t} \left(\boldsymbol{\beta}_j' \boldsymbol{\chi}_t + \boldsymbol{\gamma}_j' \boldsymbol{I}_t \right) - \left(n_{j0t} + n_{j1t} \right) \ln \left\{ 1 + \exp\left(\boldsymbol{\beta}_j' \boldsymbol{\chi}_t + \boldsymbol{\gamma}_j' \boldsymbol{I}_t \right) \right\} \right]$$

$$= \ln L_0 + \ln L_1 = l_0 + l_1 .$$

The kth element of vector $\boldsymbol{\chi}_t$ used in the computation of Equation (3.4) is the partial derivative of l_j, $j = 0, 1$, with respect to γ_{jk} at $\boldsymbol{\gamma}_j = 0$ and $\boldsymbol{\beta}_j = \hat{\boldsymbol{\beta}}_j$,

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

where, O_{jk} and E_{jk} are the observed and expected numbers of responses in the kth 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, $\boldsymbol{V}_j$, $j = 0, 1$ are as follows:

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

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

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

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

$$(\chi_{t1}, ..., \chi_{tp}) \in R_{k'}, \quad \hat{p}_{jt} = \exp\left(\hat{\boldsymbol{\beta}}_j' \boldsymbol{\chi}_t\right) \Big/ \left\{1 + \exp\left(\hat{\boldsymbol{\beta}}_j' \boldsymbol{\chi}_t\right)\right\}.$$

4 Extension of the model for higher-order Markov chains

Consider the *n*th-order Markov model for times, $t - n, t - (n - 1), ..., t - 1$, and *t*, with transition matrix, ***P*,** and its components:

$$p_{r \cdots sjt} = \Pr\left(Y_t = 1 \mid Y_{t-\mathrm{n}} = r, ..., Y_{t-2} = s, Y_{t-1} = j\right); j, s, r = 0, 1, t = 1, ..., T.$$

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

$$logit\left(p_{r \cdots sjt}\right) = \boldsymbol{\beta}_{r \cdots sj}' \boldsymbol{\chi}_t, \quad p_{r \cdots sjt} = \frac{\exp\left(\boldsymbol{\beta}_{r \cdots sj}' \boldsymbol{\chi}_t\right)}{1 + \exp\left(\boldsymbol{\beta}_{r \cdots sj}' \boldsymbol{\chi}_t\right)},$$

where vector, $\boldsymbol{\chi}_t = \left(1, x_{t1}, ..., x_{tp}\right)$, contains covariates and $\boldsymbol{\beta}_{l \cdots sj} = \left(\beta_{r \cdots sj0}, \beta_{r \cdots sj1}, ..., \beta_{r \cdots sjp}\right)$ is the vector of parameters.

The likelihood function corresponding to this model is as follows:

$$L = \prod_t \prod_{j=0}^{1} \prod_{s=0}^{1} \cdots \prod_{r=0}^{1} \left(1 - p_{r \cdots sjt}\right)^{n_{r \cdots sj0i}} p_{r \cdots sjt}^{\;n_{r \cdots sj1i}}$$

The log-likelihood is:

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

$$= \sum_{r,s,j=0}^{1} \ln \mathrm{L}_{r\cdots sj},$$

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

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

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

$$\log it\left(p_{r\cdots sjt}\right) = \boldsymbol{\beta}_{r\cdots sj}^{'} \boldsymbol{\chi}_t + \boldsymbol{\gamma}_{r\cdots sj}^{'} \boldsymbol{I}_t;\ r, s, j = 0, 1. \tag{4.1}$$

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

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

where, $\boldsymbol{Z}$ is a 2^n-dimensional vector with elements of:

$$\boldsymbol{Z}_{r\cdots sj}^{'} = \left({\partial l}/{\partial \gamma_{r\cdots sj1}}, \ldots, {\partial l}/{\partial \gamma_{r\cdots sjG}} \right),\ r, s, j = 0, 1.$$

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

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

where,

$$A_{r\cdots sjkk^{'}} = -\partial^2 l / \partial \gamma_{r\cdots sjk}\, \partial \gamma_{r\cdots sjk^{'}} \quad \left(k, k^{'} = 1, \ldots, G\right),$$

$$B_{r\cdots sjkk^{'}} = -\partial^2 l / \partial \gamma_{r\cdots sjk}\, \partial \beta_{r\cdots sjk^{'}} \quad \left(k = 1, \ldots, G; k^{'} = 0, \ldots, p\right),$$

$$C_{r\cdots sjkk'} = -\partial^2 l / \partial \beta_{r\cdots sjk}\, \partial \beta_{r\cdots sjk'} \quad \left(k, k^{'} = 0, \ldots, p\right), \left(r, s, j = 0, 1\right).$$

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

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

$$= \sum_{r,s,j=0}^{1} \ln \mathrm{L}_{r\cdots sj} = \sum_{r,s,j=0}^{1} l_{r\cdots sj}$$

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

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

5 Application

We applied the proposed test on the Health and Retirement Study (HRS) data to demonstrate its application. This is a longitudinal household survey dataset 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 8 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 $\left(\chi_{t1}, \ldots, \chi_{tp}\right)$ 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 by using $\Sigma_{ij}\left(f_{ij} - f_i\, p_{ij}\right)^2 / \left(f_i\, p_{ij}\right)$, 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 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 Tsiatis' statistics.

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

	Transition time		t	
First order		$t-1$	0	1
		0	3951	1455
		1	1930	257
Second order	$t-2$	$t-1$	t	
	0	0	3473	1396
	1	0	680	221
	0	1	269	29
	1	1	568	151

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.

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

First-order				
Transition type	Covariates	Estimated value	s.e.	p-value
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				
Transition type	Covariate	Estimated value	s.e.	p-value
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)	

6 Simulation

Data generated by the techniques provided by Ghosh and Mukerjee (2001), and Leisch et al. (1998), were 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 $\Sigma_{ij}\left(f_{ij}-f_i\, p_{ij}\right)^2/\left(f_i\, p_{ij}\right)$, 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, $H_0: \gamma_{j1} = \ldots = \gamma_{jG} = 0$, 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 4.8% to 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, $H_0: \gamma_{r \cdots sj1} = \ldots = \gamma_{r \cdots sjG} = 0$, in 500 times of simulations 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 1,000 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 Tsiatis' test was 3.6% to 6.2% for models 1 to 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 Tsiatis' 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.

Table 3: 500 Simulations for Obtaining the Estimates of Associations Based on the Proposed First-order Models

		Model 1 - size 250		Model 2 - size 500		Model 3 - size 1000		Model 4 - size 250		Model 5 - size 500	
Transition type											
00		61		121		243		50		100	
01		65		129		257		50		100	
10		14		29		58		75		150	
11		110		221		442		75		150	
Correlation of response variables		0.4		0.4		0.4		0		0	
Estimates of Parameters		estimate	p-value	estimate	p-value	estimate	p-value	estimate	p-value	estimate	p-value
0 to 1	Constant	-0.868	0.115	-0.843	0.025	-0.830	0.002	-0.916	0.130	-0.877	0.040
	V1	1.422	0.011	1.386	0.000	1.360	0.000	0.972	0.143	0.942	0.043
	V2	0.301	0.438	0.290	0.383	0.287	0.304	0.341	0.396	0.357	0.313
	V3	0.093	0.482	0.109	0.469	0.101	0.464	0.517	0.321	0.469	0.237
	V4	0.305	0.423	0.262	0.380	0.259	0.293	0.759	0.263	0.751	0.126
1 to 1	Consta	-	0.19	-	0.08	-	0.01	1.233	0.04	1.200	0.00

	nt	1.181	5	1.114	4	1.044	9		5		5
	V1	-1.776	0.059	-1.696	0.005	-1.685	0.000	-0.914	0.072	-0.907	0.012
	V2	-0.365	0.444	-0.257	0.453	-0.250	0.418	-0.364	0.352	-0.350	0.273
	V3	-0.013	0.483	-0.037	0.483	-0.061	0.468	-0.496	0.292	-0.491	0.161
	V4	-0.252	0.452	-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	
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		474		470	
Proportion of rejection of H_0		40/500		33/500		24/500		26/500		30/500	

Table 3: Continued

		Model 6 - size 1000		Model 7 - size 250		Model 8 - size 500		Model 9 - size 1000	
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
0 to 1	Constant	-0.875	0.003	0.485	0.314	0.489	0.202	0.492	0.081
	V1	0.926	0.004	1.797	0.028	1.707	0.002	1.688	0.000
	V2	0.351	0.215	2.183	0.036	1.961	0.002	1.931	0.000
	V3	0.470	0.112	0.709	0.274	0.666	0.172	0.635	0.066
	V4	0.743	0.028	-0.715	0.269	-0.666	0.166	-0.653	0.063
1 to 1	Constant	1.190	0.000	2.463	0.003	2.315	0.000	2.291	0.000
	V1	-0.885	0.000	-1.731	0.013	-1.635	0.001	-1.613	0.000
	V2	-0.342	0.154	-1.621	0.011	-1.546	0.000	-1.529	0.000
	V3	-0.491	0.060	-0.645	0.242	-0.605	0.153	-0.611	0.042
	V4	-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	0.934	0.011	0.923
No. of tests accepting H_0		500		500		500		500	
Proposed test		8.167	0.489	8.383	0.467	8.439	0.468	8.448	0.463
No. of tests accepting H_0		472		471		475		473	
Proportion of rejection of H_0		28/500		29/500		25/500		27/500	

Table 4: 500 Simulations for Obtaining the Estimates of Associations Based on the Proposed Second-order Models

	Model 1 - size 250	Model 2 - size 500	Model 3 -size 1000
Transition			
000	67	132	262
001	15	31	61
100	22	44	88
101	22	43	88
010	22	43	87
011	22	44	88
110	15	31	62
111	65	132	264

Correlation of response variables		0.4		0.4		0.4	
Estimates of parameters		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	

Table 4: Continued

	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
011	31	62	125
110	31	63	125
111	32	62	125

Correlation of response variables		0		0		0	
Estimates of parameters		Estimate	p-value	estimate	p-value	estimate	p-value
0→0→1	Constant	-0.771	0.266	-0.704	0.144	-0.724	0.034
	V1	0.615	0.343	0.577	0.260	0.581	0.125
	V2	0.912	0.288	0.847	0.163	0.838	0.053
	V3	0.412	0.400	0.378	0.362	0.406	0.240
	V4	0.381	0.441	0.320	0.422	0.317	0.373
1→0→1	Constant	-1.083	0.229	-0.976	0.111	-0.964	0.029
	V1	0.644	0.353	0.547	0.266	0.572	0.123
	V2	0.879	0.259	0.806	0.144	0.788	0.045
	V3	0.474	0.409	0.436	0.327	0.429	0.228
	V4	0.326	0.448	0.321	0.389	0.296	0.337
0→1→1	Constant	1.013	0.239	1.004	0.112	0.931	0.033
	V1	-0.580	0.359	-0.639	0.222	-0.557	0.133
	V2	-0.942	0.251	-0.795	0.177	-0.794	0.049
	V3	-0.434	0.409	-0.427	0.339	-0.401	0.251
	V4	-0.310	0.454	-0.306	0.419	-0.278	0.370
1→1→1	Constant	1.291	0.254	1.270	0.101	1.241	0.025
	V1	-0.611	0.377	-0.605	0.264	-0.597	0.146
	V2	-0.849	0.253	-0.789	0.130	-0.778	0.038
	V3	-0.450	0.434	-0.440	0.360	-0.429	0.261
	V4	-0.330	0.437	-0.300	0.421	-0.301	0.337
Billingsley		3.38E-04	1.000	7.01E-05	1.000	2.14E-05	1.000
No. of tests accepting H_0		500		500		500	
Proposed test		17.477	0.425	16.562	0.464	16.517	0.473
No. of tests accepting H_0		559		476		473	
Proportion of rejection of H_0		41/500		24/500		27/500	

Table 4: Continued

		Model 7 - size 250		Model 8 - size 500		Model 9 -size 1000	
Transition							
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		-0.4		-0.4		-0.4	
Estimates of parameters		Estimate	p-value	estimate	p-value	estimate	p-value
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
	V4	-0.091	0.498	-0.085	0.474	-0.092	0.479
1→0→1	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
	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
0→1→1	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
	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
1→1→1	Constant	0.326	0.484	0.183	0.504	0.185	0.460
	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
Billingsley		9.51E-03	1.000	3.86E-03	1.000	1.12E-03	1.000
No. of tests accepting H_0		500		500		500	
Proposed test		15.810	0.508	17.039	0.442	16.388	0.478
No. of tests accepting H_0		479		469		473	
Proportion of rejection of H_0		21/500		31/500		27/500	

7 Conclusion

An extension of Tsiatis' 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.

Acknowledgments

We 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. We are also 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.

References

[1] Anastasios A. Tsiatis, A Note on a Goodness-of-Fit Test for The Logistic Regression Model, Biometrika, 67 (1980) 250- 251.

[2] C. R. Rao, *Linear Statistical Inference and its Applications*, 2nd edition, New York: Wiley, (1973).

[3] G. McQueen and S. Thorley, *Are Stock Returns Predictable? A Test Using Markov Chains*, The Journal of Finance, 46 (1991) 239-263.

[4] G. Y. Yi, W. He, and H. Liang, *Analysis of correlated binary data under partially linear single-index logistic models*, Journal of Multivariate Analysis, 100 (2009) 278-290.

[5] Health and Retirement Study (HRS), Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG09740). Waves [1-8], Year [1992-2006] (2009) [Online], [Accessed 2009]. Available from World Wide Web: http://hrsonline.isr.umich.edu/data/index.html

[6] L.R. Muenz, L.V. Rubinstein, *Markov models for covariate dependence of binary sequences*, Biometrics 41 (1985) 91–101.

[7] M. A. Islam, R. I. Chowdhury, *A higher order Markov model for analyzing covariate dependence*, Applied Mathematical, 30 (2006) 477-488.

[8] M. A. Islam, R. I. Chowdhury, and L. Briollais. *A bivariate binary model for testing dependence in outcomes*. Bull. Malays. Math. Sci. Soc.(2) 35.4 (2012) 845-858.

[9] M. Rahman Shafiqur and M. A. Islam, *Markov structure based logistic regression for repeated measures: An application to diabetes mellitus data*, Statistical Methodology, 4 (2007) 448-460.

[10] M. S. Bartlett, *The frequency goodness of fit test for probability chains*, Proc. Camb. Phil. Soc., 47 (1951) 86.

[11] P. Billingsley, *Statistical Methods in Markov Chains*, The Annals of Mathematical Statistics, 32 (1961) 12-40.

[12] P. S. Albert, *A Markov Model for Sequences of Ordinal Data from a Relapsing-Remitting Disease*, International Biometric Society, 50 (1994) 51-60.

[13] R. I. Chowdhury, M. A. Islam, M. A. Shah, N. Al-Enezi, *A Computer Program to Estimate the Parameters of Covariate Dependence Higher Order Markov Model*, Computer Methods and Program in Biomedicine, 77 (2005) 175-181.

[14] Sirdari, M.Z., Islam, M.A. and Awang, N., *A stationarity test on Markov chain models based on marginal distribution*, Statistical Methodology, 11 (2013) 68-76.

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

[16] M. A. Islam, R. I. Chowdhury, and L. Briollais. *A bivariate binary model for testing dependence in outcomes*. Bull. Malays. Math. Sci. Soc.(2) 35.4 (2012) 845-858.

In this paper, an important test procedure arising from repeated measurements of discrete data has 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.

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 modelling, 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.