Open Journal of Statistics, 2016, 6, 469-481
Published Online June 2016 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2016.63042

Transition Logic Regression Method to Identify Interactions in Binary Longitudinal Data

Parvin Sarbakhsh[1], Yadollah Mehrabi[2*], Jeanine J. Houwing-Duistermaat[3], Farid Zayeri[4], Maryam Sadat Daneshpour[5]

[1]Department of Statistics and Epidemiology, School of Public Health, Tabriz University of Medical Sciences, Tabriz, Iran
[2]Department of Epidemiology, School of Public Health, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[3]Department of Medical Statistics and Bioinformatics, Leiden University Medical Centre, Leiden, Netherlands
[4]Department of Biostatistics, School of Paramedicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[5]Cellular and Molecular Endocrine Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
Email: p.sarbakhsh@gmail.com, *mehrabi@sbmu.ac.ir, j.j.houwing@lumc.nl, fzayeri@yahoo.com, daneshpour1388@gmail.com

Received 20 October 2015; accepted 19 June 2016; published 22 June 2016

Abstract

Logic regression is an adaptive regression method which searches for Boolean (logic) combinations of binary variables that best explain the variability in the outcome, and thus, it reveals interaction effects which are associated with the response. In this study, we extended logic regression to longitudinal data with binary response and proposed "Transition Logic Regression Method" to find interactions related to response. In this method, interaction effects over time were found by Annealing Algorithm with AIC (Akaike Information Criterion) as the score function of the model. Also, first and second orders Markov dependence were allowed to capture the correlation among successive observations of the same individual in longitudinal binary response. Performance of the method was evaluated with simulation study in various conditions. Proposed method was used to find interactions of SNPs and other risk factors related to low HDL over time in data of 329 participants of longitudinal TLGS study.

Keywords

Logic Regression, Longitudinal Data, Transition Model, Interaction, TLGS Study, Low HDL, SNP

*Corresponding author.

How to cite this paper: Sarbakhsh, P., Mehrabi, Y., Houwing-Duistermaat, J.J., Zayeri, F. and Daneshpour, M.S. (2016) Transition Logic Regression Method to Identify Interactions in Binary Longitudinal Data. *Open Journal of Statistics*, **6**, 469-481.
http://dx.doi.org/10.4236/ojs.2016.63042

1. Introduction

Regression analysis is an important tool in evaluating the functional relationship between dependent variable, and a set of independent variables. On most issues, regression models can only relate the main effects of predictor variables to the response variable and evaluation of interaction effects cannot be exceeded of two-way or at most three-way, due to complexity of such interactions.

In order to consider such interactions in the regression models, some combinations of explanatory variables can be constructed and these combinations can be used as new predictors instead of using individual variables.

"Logic Regression" is a type of generalized regression and classification method based on logic combinations of binary variables which can make Boolean combinations of original binary explanatory variables in order to reveal interactions [1]. Logic regression is different from logistic regression with "logit" link function that is a member of generalized linear model family for modeling response variables with binomial distribution. Although we can evaluate interactions using logistic regression, these interactions need to be known in advance, and used as input variables in the model. By contrast, Logic Regression is applicable for any type of response, as long as the predictors are binary. Interactions of interest need not be known in advance, quite the contrary, the detection of important variable interactions is the main aim of logic regression [2]. Logic regression is introduced and used for case control or cohort studies with independent observations [2].

Furthermore, some extensions have been performed to this model in several ways. Namely, Multinomial Logic Regression has been developed for multinomial categorical responses [2]. Trio Logic Regression with conditional Logic Regression model has been proposed to analyze data of case parents trios [3]. Monte Carlo Logic Regression has been developed to generate a list of predictors related to the response [4]. Logic FS has been introduced and used to identify different Logic Regressions associated with response [5]. Genetic programming for association studies [6] has been proposed for classification settings, and uses genetic programming as search algorithm.

On the other hands, a longitudinal study is defined as an investigation where subject's responses are recorded at multiple follow-up times. A longitudinal study yields "repeated measurements" on each subject. In compare to cross sectional studies, longitudinal studies have some benefits such as measurement of individual change in outcomes, separation of time effects, and control for cohort effects [7].

Like other kind of regression models, interactions among predictors are important in modelling of longitudinal data. In addition, one of the goals of longitudinal studies is to examine whether the relationship between the response and the predictors changes over time. In other words, if there is any interaction between variables and time or not. It seems that logic regression theory can be used to assess interactions in modeling of longitudinal data. To find such time dependent interactions in quantitative longitudinal response, recently, "logic mixed model", based on linear mixed model, has been proposed and used to assess the interactions of SNP associated with longitudinal quantitative cholesterol level [8], but Logic Regression has not been developed for analysis of correlated binary observations of longitudinal studies up to now.

So, due to the importance of the interactions related to such responses, in this paper we proposed "Transition Logic Regression" model as an extension of logic regression to detect and assess higher order interactions over time in longitudinal data with binary response. Furthermore, we carried out a simulation study to evaluate the performance of our model in different settings and compare it with standard model. In addition, as an application, we assessed effects of some SNPs and other risk factors on having low level of HDL over time using our proposed Transition Logic Regression model.

The present paper was initially motivated by the SNP dataset with potential important interactions among SNPs related to binary longitudinal response.

2. Method

2.1. Logic Regression

Logic Regression is a generalized regression and classification method that enables identification of interactions by using Boolean combinations as new independent variables of the original binary variables. We try to find Boolean statements involving the binary predictors that enhance the prediction for the response. These Boolean combinations are logic expressions such as $L_1 = (X_1 \wedge X_3) \vee (X_5 \wedge X_7^c)$. It means that if the response is binary as well (which is not required in general), we attempt to find decision rules such as "if X_1 and X_3 are true",

or " X_5 but not X_7 are true", then the response is more likely to be in class 0.

Let $X_1,\cdots,X_k$ be binary predictors, Y be a response variable and $Z_1,\cdots,Z_p$ be quantitative covariates, Logic Regression models are of the form: $g\left(E\left(Y\right)\right)=\beta_0+\sum_{i=1}^{p}\gamma_i Z_i+\sum_{j=1}^{t}\beta_j L_j$

where g is a link function for response and L_j is a Boolean combination of the binary predictors X_i.

Logic regression is an adaptive algorithm which for a given model selects those L_j that minimize the score function of the model. Logic Regression framework includes many forms of regression (such as linear and logistic regression, Cox proportional hazards model). For every model type a score function is defined indicating the "quality" of the model. In general, any type of model can be considered, as long as a scoring function (such as a deviance or likelihood) is defined [2].

2.2. Simulated Annealing for Logic Regression

The number of logic expressions that can be built from a given set of binary predictors is huge, and there is no straight method to enlist all logic terms that yield different score. So, it is infeasible to do an exhaustive assessment of all different logic terms and select the best model. In order to solve this problem in Logic Regression, a simulated annealing as a stochastic search algorithm is used to search for the best logic combinations and estimate the β_j [1].

There are some permissible moves in logic regression theory such as alternating a predictor, alternating an operator, deleting a predictor and so on, which called permissible moves. These moves are used in Annealing algorithm to generate new logic expressions in the search for the best logic regression model according to a score function. For more information about permissible moves see [1]. In each iteration of the simulated annealing algorithm, a new logic term is proposed by randomly executing a move from the set of permissible move and so related new Logic Regression model is fitted. The acceptance probability for the new logic term is based on the score function of the new and current models, and a simulated annealing parameter called temperature [2].

2.3. Transition Model: Marginal Modelling of Binary Longitudinal Data Using Markov Chains

In order to extend Logic Regression to longitudinal study, we considered one kind of transition model for binary longitudinal data introduced by Gonçalves [9]. This model is a marginal modelling of binary longitudinal data using Markov chains. Below this model is briefly described.

For notation, $Y_{it}\in\{0,1\}$ is binary response variables of individual i ($i=1,\cdots,n$) at time t $\left(t=1,\cdots,T_i\right)$, with mean $P\left(Y_{it}=1\right)=\theta_{it}$. For each subject at each time, let $\boldsymbol{X}_{it}$ be a set of p covariates that first column of its can be a vector of ones to consider intercept term. Logistic regression model that marginally connects the probability distribution of the response and auxiliary variables is:

$$\text{logit}\, P\left(Y_{it}=1\right)=\boldsymbol{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}$$

where β is a p vector of unknown parameters. To take into account the correlation among successive observations of the same individual, the model considers a Markovian type of first order (ψ_1) or of second order (ψ_2) dependence structure. For the sake of simplicity, the subject subscript i was ignored, since individuals are assumed to be independent from each other. In the first order binary Markov chain model, the joint distribution $\left(Y_1,\cdots,Y_T\right)$ are determined by the distribution of Y_1 and a set of conditional probabilities:

$$p_j=P\left\{Y_t=1\,|\,Y_{t-1}=j\right\},\left(t=2,\cdots,T\right),\left(j=0,1\right)$$

For a pair of successive observations $\left(Y_t,Y_{t-1}\right)$ with known marginal distribution of Y_{t-1}, p_j is chosen so that $\theta_t=E\left(Y_t\right)$ is already assigned.

In order to analyze the binary data, the quantity odds ratio is the preferred measure of dependence between observations:

$$OR\left(Y_t,Y_{t-1}\right)=\psi_1=\frac{P\left\{Y_{t-1}=Y_t=1\right\}P\left\{Y_{t-1}=Y_t=0\right\}}{P\left\{Y_{t-1}=0,Y_t=1\right\}P\left\{Y_{t-1}=1,Y_t=0\right\}}=\frac{p_1/\left(1-p_1\right)}{p_0/\left(1-p_0\right)}$$

After solving following equations with respect to p_0 and p_1:

$$\psi_1 = \frac{p_1/(1-p_1)}{p_0/(1-p_0)},\ \theta_t = \theta_{t-1}p_1 + (1-\theta_{t-1})\,p_0$$

It yields:

$$p_j = \frac{(2j-1)\left[1-\delta_0+(\psi_1-1)\theta_{t-1}\right]+(\psi_1-1)\theta_t}{2(\psi_1-1)\left[1-j+(2j-1)\theta_{t-1}\right]},\ j=(0,1)$$

where $\delta_0^2 = 1+(\psi_1-1)\left\{\psi_1(\theta_t-\theta_{t-1})^2-(\theta_t+\theta_{t-1})^2+2(\theta_t+\theta_{t-1})\right\}$.

If $\psi_1 = 1$, the variables are independent and $p_j = \theta_t$. Similarly, in the second order binary Markov chain model, for (Y_{t-2}, Y_{t-1}, Y_t) transition probabilities are:

$$p_{hj} = P\{Y_t = 1 \mid Y_{t-2} = h, Y_{t-1} = j\}, (t = 2,\cdots,T),\ h, j = 0,1$$

First and second order dependence are:

$$OR(Y_{t-1}, Y_{t-2}) = \psi_1 = OR(Y_{t-1}, Y_t) \tag{1}$$

$$OR(Y_{t-2}, Y_t \mid Y_{t-1} = 0) = \psi_2 = OR(Y_{t-2}, Y_t \mid Y_{t-1} = 1) \tag{2}$$

p_{hj} can be calculated using these equations:

$$p'_j = p\{Y_t = 1 \mid Y_{t-1} = j\},\ p''_j = p\{Y_{t-1} = 1 \mid Y_{t-2} = j\}$$

$$\frac{p'_1(1-p'_0)}{p'_0(1-p'_1)} = \psi_1 = \frac{p''_1(1-p''_0)}{p''_0(1-p''_1)}$$

$$\frac{p_{10}(1-p_{00})}{p_{00}(1-p_{10})} = \psi_2 = \frac{p_{11}(1-p_{01})}{p_{01}(1-p_{11})}$$

$$\theta_t = p_{11}p''_1\theta_{t-2} + p_{10}(1-p''_1)\theta_{t-2} + p_{01}p''_0(1-\theta_{t-2}) + p_{00}(1-p''_0)(1-\theta_{t-2})$$

$$p'_1\theta_{t-2} = p_{11}p''_1\theta_{t-2} + p_{01}p''_0(1-\theta_{t-2})$$

Likelihood inference is performed based on sample of n subjects who are assumed to be independent from each other. If y_{it} is observation of subject i ($i = 1,\cdots,n$) at time t ($t = 1,\cdots,T$), the contribution of subject i with all observations y_i. to the log likelihood function for the parameters (β, λ) is:

$$\begin{aligned}&\log\text{likelihood}(\boldsymbol{\beta},\boldsymbol{\lambda} \mid \boldsymbol{y}_i)\\ &= \left[y_{i1}\text{logit}(\theta_1) + \log(1-\theta_1)\right] + \left[y_{i2}\text{logit}\left(p'_{y_{i1}}\right) + \log\left(1-p'_{y_{i1}}\right)\right]\\ &\quad + \sum_{t=3}^{T}\left[y_{it}\text{logit}\left(p_{y_{it-2},y_{it-1}}\right) + \log\left(1-p_{y_{it-2},y_{it-1}}\right)\right]\end{aligned}$$

where $\boldsymbol{\lambda} = (\log\psi_1, \log\psi_2)$.

Clearly, the likelihood function for the entire sample is obtained by calculating the sum of the likelihood of all subjects [9]: $\log\text{likelihood} = \sum_{i=1}^{n}\log\text{likelihood}(\boldsymbol{\beta},\boldsymbol{\lambda} \mid \boldsymbol{y}_i)$.

AIC statistic for the model is calculated as:

$$\text{AIC} = -2\log\text{likelihood} + 2q \tag{3}$$

where q equals the number of parameters in the model.

2.4. Our Proposed Method: Transition Logic Regression

In this paper, mentioned first and second order Markov chain transition model with AIC (Equation (3)) as a score function of the model, was used to develop Logic Regression to longitudinal data. Therefore, "Transition Logic Regression" was defined as: $\text{logit}\,P(Y_{it}=1) = \boldsymbol{Z}_{it}^{\mathrm{T}}\boldsymbol{\gamma} + \boldsymbol{L}_{it}^{\mathrm{T}}\boldsymbol{\beta}$ which $Y_{it} \in \{0,1\}$ is binary response variables

of individual i ($i=1,\cdots,n$) at time t $(t=1,\cdots,T_i)$, with mean $P(Y_{it}=1)=\theta_{it}$ and $\boldsymbol{Z}_{it}$ is vector of quantitative covariates and $\boldsymbol{L}_{it}$ is vector of Boolean expression from binary predictors X_{it}. $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are vectors of unknown parameters. To take into account the correlation among successive observations of the same individual, the model considers a Markovian type of first order (ψ_1) or of second order (ψ_2) dependence structure (Equations (1) and (2)).

Searching to find best $\boldsymbol{L}_{it}$ so that the fitted model has low AIC, was done using Annealing algorithm. Therefore, Annealing algorithm searched for Boolean combinations which according to the AIC statistic had the lowest score and therefore had the best fitting in Transition Logic Regression model. This extension allows for the fit of a Transition Logic Regression model. The program of Transition Logic Regression was written in FORTRAN 77 and added to "LogicReg" package [1]. Modified "LogicReg" package was recompiled an installed in R(2.15.3) to analyze data.

3. Simulation Study

Simulation study was done to assess the performance of proposed model and to compare it with the standard model. Data was produced from binomial distribution with first order Markov chain dependence structure for three time points.

Given specific sample size, for each sample in time *t*, ten covariates were simulated from Bernoulli (5):

$$X_{pt} \sim \text{bernoulli}(0.5)(p=1,\cdots,10),(t=1,2,3)$$

The simulated model assumed $L_t = X_{1t} \vee X_{2t}$ as the interaction effect between predictors at time t. For each sample in each time t, three repeated measurements were constructed as the response variable Y_t each with a predetermined probability of success θ_t related to the interaction L_t via logit link function:

$$\text{logit}\, P(Y_t=1) = \log\left(\frac{\theta_t}{1-\theta_t}\right) = L_t^{\mathrm{T}}\beta$$

$$\theta_t = 1/\left(1+\exp\left(-\beta * L_t\right)\right)$$

Starting with the first response, y_1 was produced from Bernoulli distribution with mean θ_1 Transition probabilities are:

$$p_1 = p(y_2=1 \mid y_1=1)$$

$$p_0 = p(y_2=1 \mid y_1=0)$$

Respect to our desired values of θ_t and ψ_1, these first order transition probabilities were calculated. So, if y_1 equals to one, y_2 produced from Bernoulli distributed with probability of p_1 else if y_1 equals to zero, y_2 was simulated from Bernoulli distributed with mean p_0.

In order to produce y_3 under desired consideration, we calculated following transition probabilities:

y_3 under desired consideration, we calculated below transition probabilities:

$$p_1' = p(y_3=1 \mid y_2=1)$$

$$p_0' = p(y_3=1 \mid y_2=0)$$

To simulate y_3, if y_2 equals to one, y_3 is simulated from Bernoulli distributed with probability p_1' and if y_2 equals to zero, y_3 is produced from Bernoulli distribution with mean p_0'.

$$\begin{cases} y_2=1 \Rightarrow y_3 \sim \text{bernoulli}(p_1') \\ y_2=0 \Rightarrow y_3 \sim \text{bernoulli}(p_0') \end{cases}$$

Simulation study was done for various sample sizes (number of cases: 50, 200, 500, 1000), first order Markov chain dependences $(\psi_1 = 0.2, 0.5, 1, 2, 5)$, and coefficients of the interaction term $(\beta = 0, 0.5, 1.5, 3)$.

With respect to simulated interaction term, we considered all covariates as the search space and one combination with two variables as the model size in annealing algorithm setting. For this simulation study, 500 datasets were generated for each condition.

Percentage of identification of exact simulated interaction was considered as quality of performance of the Transition Logic Regression model. Also, AIC of Transition Logic Regression was compared with AIC of Transition model as the standard model which only includes all ten covariates as the main effects in the model.

In addition, MSE and 95% empirical confidence interval of estimators in models that could identify interaction truly were calculated. Lower bound of empirical confidence intervals is 0.025th quantile and upper bound is 0.975th quantile of estimated values of parameters.

The results of simulation study are shown in **Tables 1-4**. According to these tables, as expected with increasing sample size and coefficient of interaction term, the rate of identification of true interaction increases. For example, in $n = 200$ and $\beta = 3$ method was able to find true interaction term in all 500 data sets. The value of the first order dependence did not have considerable effect on the performance of the method.

The same holds, MSE and confidence intervals of estimations get better with increasing of sample size. In small sample sizes, amount of coefficient of interaction and first order dependence have effect on MSE of ψ_1 so that in strong interaction effect or strong first order dependence, MSE of ψ_1 is large.

Maximum type I error was 0.01 that method had found $L = X_1 \vee X_2$ as interaction effect when there was not such interaction in data $(\beta = 0)$.

4. Application of Proposed Model on TLGS Data

Interactions usually play an important role in SNP (Single-nucleotide polymorphism) association studies. High order interactions of SNPs are supposed to explain the differences between low- and high-risk groups [10]. In addition to the main effects of SNPs, their interactions are assumed to be responsible for low HDL. SNPs interactions can be time-dependent. So, our aim of this study was investigation SNPs interactions related to low HDL over time. Subjects in this study were selected from among participants of the Tehran Lipid and Glucose Study (TLGS). TLGS is a prospective study to determine the risk factors and outcomes of non-communicable disease [11]. The structure of this study includes some major components. The TLGS design has been explained elsewhere [12]. Longitudinal data from the three phases of the TLGS study was analyzed to assess the association between the some related polymorphisms and other risk factors with low levels of HDL over time. In order to assess this association, Transition Logic Regression models with first and second order Markov chain were fitted.

First order Markov chain Transition Logic Regression model with three tree logic (Boolean combination) and 8 leaves (predictor variables) was fitted.

A total of 329 subjects (127 (38.6%) men and 202 (61.4%) women) who were present in phase I, II, III of TLGS study with age ≥20 years and without any missing value in evaluated variables were randomly selected and included in the current study.

Low HDL-C level was defined as <40 mg/dL for men and <50 mg/dL for women. High waist circumference (WC) was defined as WC ≥95 cm for Iranian men and women [13]. High triglyceride (TG) level was defined as TG ≥150 mg/dL, subjects who had blood pressure (BP) ≥130/85 mmHg or used anti-hypertension drug, and subjects with fasting blood sugar (FBS) ≥110 mg/dL or users of anti-diabetic drugs were considered as high BP and high FBS respectively [14] [15]. Subjects who smoke daily or occasionally were considered as smokers. Phase of study was considered as time.

Table 5 pictures the summary of demographic characteristic and clinical and lipid profiles of these subjects in three phases of study. Highest prevalence of having low HDL (79.3%) was seen in phase 2 of study.

The polymorphisms of ApoA1M1, ApoA1M2, ApoB, ApoAIV, ApoCIII, ABCA1, SRB1 and ApoE genes that have been shown to be associated with HDL-C disorder [16]-[20] were investigated. Allele frequencies given in **Table 6** show genotype distributions. The +/+ genotype of Apo A1M2 gene had the highest prevalence (91.2%) and TT genotype of Apo AIV gene had the lowest frequency (0.3%).

Each SNP was considered as a random variable taking values 0, 1, and 2 corresponding to the nucleotide pairs. We coded each of these variables into two dummy binary variables corresponding to a dominant and a recessive effect. By this approach, we generated 2p binary predictors out of p SNPs to perform interaction terms for Logic Regression [1].

The results of Transition Logic Regression with first order Markov chain show that subjects with high triglyceride and high waist circumstance have an odds ratio of 2.29 to have low level of HDL. Also, (being in phase 2 and ((carrier of the minor allele of ApoA1M1) or (being homozygous for the common allele of ApoCIII))) was

Table 1. Result of simulation study with n = 50.

$n = 50$	$\beta = 0$					$\beta = 0.5$					$\beta = 1.5$					$\beta = 3$				
	$\psi_1 = 0.2$	$\psi_1 = 0.5$	$\psi_1 = 1$	$\psi_1 = 2$	$\psi_1 = 5$	$\psi_1 = 0.2$	$\psi_1 = 0.5$	$\psi_1 = 1$	$\psi_1 = 2$	$\psi_1 = 5$	$\psi_1 = 0.2$	$\psi_1 = 0.5$	$\psi_1 = 1$	$\psi_1 = 2$	$\psi_1 = 5$	$\psi_1 = 0.2$	$\psi_1 = 0.5$	$\psi_1 = 1$	$\psi_1 = 2$	$\psi_1 = 5$
*AIC of transition model with only main effects	204.27	216.36	218.61	216.83	204.84	20.03	21.23	212.91	21.36	199.65	167.54	172.58	173.81	172.21	165.8	107.29	109.34	109.75	109.51	107.24
**AIC of transition logic regression model	194.14	20.28	†-	206.35	-	186.49	194.48	196.83	195.14	181.39	153.11	158.35	159.32	157.16	151.31	94.88	96.7	97.15	96.55	94.47
Percentage of identification of true interaction	1	0.5	0	0.5	0	5	4.4	3.2	5.6	6	64	6.6	6.2	61.4	67	98.4	97.6	96.6	97.4	98
** $\hat{\psi}_1$ (confidence interval)	0.22 (0.18, 0.26)	1.49 (1.5, 1.5)	-	1.79 (1.79, 1.79)	-	0.15 (0.06, 0.38)	0.51 (0.3, 2.23)	0.9 (0.43, 2.23)	1.7 (0.82, 3.16)	6.62 (3.76, 11.21)	0.1 (0, 0.57)	0.41 (0.11, 1.38)	0.97 (0.21, 3.02)	1.99 (0.53, 6.87)	5.31 (1.74, 2.29)	0.04 (0, 2.01)	0.2 (0, 4.17)	0.61 (0, 11.62)	1.52 (0.04, 31.06)	4.39 (0.26, 42.88)
** MSE$\hat{\psi}_1$	0.00	0.99	-	0.05	-	0.01	0.1	0.55	0.55	1.72	0.02	0.1	1.63	4.02	22.42	0.54	26.85	28.42	48.17	55.1
** $\hat{\beta}$ (confidence interval)	0.95 (0.79, 1.11)	1.59 (1.59, 1.59)	-	1.06 (1.06, 1.06)	-	1.13 (0.81, 1.37)	1.17 (1.01, 1.41)	1.25 (1.19, 1.34)	1.19 (1.01, 1.5)	1.05 (0.89, 1.25)	1.67 (1.12, 2.32)	1.68 (1.13, 2.3)	1.67 (1.04, 2.32)	1.8 (1.17, 2.56)	1.77 (1.23, 2.53)	3.1 (1.99, 4.46)	3.1 (1.98, 4.57)	3.15 (2.06, 4.79)	3.25 (2.18, 4.76)	3.25 (2.01, 4.86)
** MSE$\hat{\beta}$	0.94	2.54	-	1.13	-	0.43	0.46	0.57	0.5	0.32	0.12	0.12	0.14	0.23	0.2	0.4	0.41	0.42	0.53	0.81

*Mean of AIC of of transition models only with main effects of covariates. **Mean of AIC, mean of $\hat{\psi}_1$, mean of $\hat{\beta}$, confidence interval and MSE of estimators in transition logic regression models that identified interaction effect correctly. †There is no model with true interaction effect to calculate indexes.

Table 2. Result of simulation study with n = 200.

$n = 200$	$\beta = 0$					$\beta = 0.5$					$\beta = 1.5$					$\beta = 3$				
	$\psi_1 = 0.2$	$\psi_1 = 0.5$	$\psi_1 = 1$	$\psi_1 = 2$	$\psi_1 = 5$	$\psi_1 = 0.2$	$\psi_1 = 0.5$	$\psi_1 = 1$	$\psi_1 = 2$	$\psi_1 = 5$	$\psi_1 = 0.2$	$\psi_1 = 0.5$	$\psi_1 = 1$	$\psi_1 = 2$	$\psi_1 = 5$	$\psi_1 = 0.2$	$\psi_1 = 0.5$	$\psi_1 = 1$	$\psi_1 = 2$	$\psi_1 = 5$
*AIC of transition model with only main effects	782.82	831.58	843.85	83.82	782.25	769.05	808.38	818.92	808.17	763.81	639.45	656.61	661.28	656.11	631.01	401.89	407.43	408.23	407.22	40.48
**AIC of transition logic regression model	†-	-	832.46	819.13	-	753.61	791.22	803.07	79.8	748.01	615.7	633.72	638.86	633.18	606.51	374.65	38.22	381.21	379.72	372.72
Percentage of identification of true interaction	0	0	1	1	0	39.8	32.8	29.6	28.6	36	99.8	99.2	99	99.8	100	100	100	100	100	100
** $\hat{\psi}_1$ (confidence interval)	-	-	0.94 (0.94, 0.94)	2.23 (2.22, 2.22)	-	0.19 (0.13, 0.31)	0.47 (0.3, 0.75)	0.91 (.57, 1.38)	1.97 (1.44, 2.91)	5.42 (3.16, 8.53)	0.18 (0.05, 0.38)	0.47 (0.21, 0.84)	1 (0.57, 1.72)	1.99 (1.09, 3.64)	5.05 (2.94, 8.74)	0.16 (0.02, 0.59)	0.47 (0.15, 1.76)	0.93 (0.35, 2.56)	1.88 (0.63, 5.54)	4.9 (1.7, 14.55)
** MSE$\hat{\psi}_1$	-	-	0	0.05	-	0	0.02	0.05	0.18	2.74	0.01	0.03	0.11	0.4	2.37	0.03	0.26	0.45	1.88	13.14
** $\hat{\beta}$ (confidence interval)	-	-	0.45 (0.45, 0.45)	0.37 (0.37, 0.37)	-	0.63 (0.42, 0.89)	0.74 (0.45, 0.94)	0.68 (0.46, 0.93)	0.7 (0.49, 0.92)	0.64 (0.43, 0.98)	1.51 (1.19, 1.97)	1.51 (1.15, 1.97)	1.52 (1.22, 1.86)	1.5 (1.12, 1.89)	1.5 (1.14, 1.86)	3.03 (2.56, 3.59)	3.04 (2.53, 3.6)	3.05 (2.64, 3.58)	3.02 (2.55, 3.63)	3.03 (2.51, 3.56)
** MSE$\hat{\beta}$	-	-	0.2	0.13	-	0.03	0.08	0.05	0.06	0.05	0.04	0.05	0.04	0.04	0.03	0.08	0.08	0.07	0.08	0.07

*Mean of AIC of of transition models only with main effects of covariates. **Mean of AIC, mean of $\hat{\psi}_1$, mean of $\hat{\beta}$, confidence interval and MSE of estimators in transition logic regression models that identified interaction effect correctly. †There is no model with true interaction effect to calculate indexes.

Table 3. Result of simulation study with n = 500.

$n = 500$	$\beta = 0$					$\beta = 0.5$					$\beta = 1.5$					$\beta = 3$				
	$\psi_1 = 0.2$	$\psi_1 = 0.5$	$\psi_1 = 1$	$\psi_1 = 2$	$\psi_1 = 5$	$\psi_1 = 0.2$	$\psi_1 = 0.5$	$\psi_1 = 1$	$\psi_1 = 2$	$\psi_1 = 5$	$\psi_1 = 0.2$	$\psi_1 = 0.5$	$\psi_1 = 1$	$\psi_1 = 2$	$\psi_1 = 5$	$\psi_1 = 0.2$	$\psi_1 = 0.5$	$\psi_1 = 1$	$\psi_1 = 2$	$\psi_1 = 5$
*AIC of transition model with only main effects	1944.78	2061.6	2091.71	2062.28	1946.41	190.5	2002.91	2029.85	2002.25	1891.22	158.92	1624.46	1637.12	1622.19	1559.65	988.56	1002.72	1007.42	1003.04	984.77
**AIC of transition logic regression model	†-	-	-	-	1963.89	1883.88	1985.07	2012.32	1985.25	1874.76	1534.82	1579.64	1592.63	1576.88	151.79	932.04	946.76	951.44	947.25	927.56
Percentage of identification of true interaction	0	0	0	0	0.5	83.8	78.4	76.8	77	85	100	100	100	100	100	100	100	100	100	100
** $\hat{\psi}_1$ (confidence interval)	-	-	-	-	4.01 (4.03, 4.03)	0.19 (0.16, 0.25)	0.51 (0.42, 0.6)	0.93 (0.78, 1.12)	1.88 (1.49, 2.34)	5.05 (3.94, 6.22)	0.19 (0.13, 0.25)	0.48 (0.37, 0.62)	0.94 (0.75, 1.41)	1.97 (1.32, 2.53)	5 (3.75, 6.56)	0.2 (0.08, 0.32)	0.52 (0.27, 0.95)	0.94 (0.45, 1.61)	2.03 (1.23, 3.32)	5.05 (3.36, 7.83)
** MSE$\hat{\psi}_1$	-	-	-	-	0.93	0	0	0.02	0.08	0.44	0	0.01	0.03	0.13	0.95	0	0.04	0.12	0.38	1.83
** $\hat{\beta}$ (confidence interval)	-	-	-	-	0.23 (0.23, 0.23)	0.53 (0.33, 0.68)	0.52 (0.4, 0.65)	0.51 (0.41, 0.61)	0.55 (0.38, 0.66)	0.53 (0.4, 0.67)	1.47 (1.31, 1.66)	1.49 (1.33, 1.64)	1.5 (1.35, 1.69)	1.52 (1.29, 1.72)	1.52 (1.27, 1.75)	3 (2.66, 3.21)	3.02 (2.72, 3.22)	3 (2.77, 3.23)	3.03 (2.66, 3.25)	3.01 (2.64, 3.21)
** MSE$\hat{\beta}$	-	-	-	-	0.05	0.01	0	0	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.03	0.03

*Mean of AIC of of transition models only with main effects of covariates. **Mean of AIC, mean of $\hat{\psi}_1$, mean of $\hat{\beta}$, confidence interval and MSE of estimators in transition logic regression models that identified interaction effect correctly. †There is no model with true interaction effect to calculate indexes.

Table 4. Result of simulation study with n = 1000.

$n = 1000$	$\beta = 0$					$\beta = 0.5$					$\beta = 1.5$					$\beta = 3$				
	$\psi_1 = 0.2$	$\psi_1 = 0.5$	$\psi_1 = 1$	$\psi_1 = 2$	$\psi_1 = 5$	$\psi_1 = 0.2$	$\psi_1 = 0.5$	$\psi_1 = 1$	$\psi_1 = 2$	$\psi_1 = 5$	$\psi_1 = 0.2$	$\psi_1 = 0.5$	$\psi_1 = 1$	$\psi_1 = 2$	$\psi_1 = 5$	$\psi_1 = 0.2$	$\psi_1 = 0.5$	$\psi_1 = 1$	$\psi_1 = 2$	$\psi_1 = 5$
*AIC of transition model with only main effects	3886.45	4116.6	417.2	4117.67	3888.98	3789.43	3993.51	4045.02	3989.87	3767.57	3145.94	3232.18	3257.62	3229.06	3103.42	1966.27	1995.88	2003.73	1994.57	1958.63
**AIC of transition logic regression model	†-	-	-	-	-	3767.07	3972.64	4024.39	3969.04	3744.18	3061.13	315.47	3176.65	3148.06	3014.96	1862.22	1892.44	190.11	1891.52	1853.8
Percentage of identification of true interaction	0.00	0.00	0.00	0.00	0.00	99.2	97.6	97.2	96.6	98.4	100	100	100	100	100	100	100	100	100	100
** $\hat{\psi}_1$ (confidence interval)	-	-	-	-	-	0.21 (0.17, 0.24)	0.51 (0.45, 0.6)	1 (0.88, 1.16)	2.01 (1.79, 2.29)	4.9 (4.26, 5.93)	0.21 (0.16, 0.28)	0.53 (0.45, 0.62)	1.03 (0.8, 1.35)	2.05 (1.69, 2.45)	5.1 (4.37, 5.69)	0.21 (0.14, 0.34)	0.51 (0.35, 0.83)	0.96 (0.57, 1.61)	1.92 (1.47, 2.87)	4.9 (3.52, 6.73)
** MSE$\hat{\psi}_1$	-	-	-	-	-	0.00	0.00	0.01	0.02	0.27	0.00	0.00	0.02	0.05	0.2	0.00	0.02	0.08	0.2	1.03
** $\hat{\beta}$ (confidence interval)	-	-	-	-	-	0.49 (0.42, 0.6)	0.5 (0.31, 0.69)	0.5 (0.32, 0.63)	0.51 (0.37, 0.64)	0.54 (0.39, 0.66)	1.5 (1.32, 1.61)	1.53 (1.3, 1.7)	1.5 (1.33, 1.67)	1.51 (1.41, 1.66)	1.49 (1.38, 1.63)	3 (2.7, 3.35)	3.01 (2.68, 3.34)	2.99 (2.71, 3.32)	3 (2.78, 3.26)	3 (2.73, 3.24)
** MSE$\hat{\beta}$	-	-	-	-	-	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.03	0.03	0.03	0.02	0.02

*Mean of AIC of of transition models only with main effects of covariates. **Mean of AIC, mean of $\hat{\psi}_1$, mean of $\hat{\beta}$, confidence interval and MSE of estimators in transition logic regression models that identified interaction effect correctly. †There is no model with true interaction effect to calculate indexes.

Table 5. Demographic characteristic and clinical and lipid profiles of subjects in phases of study.

Variables	Phase 1 (n = 329)	Phase 2 (n = 329)	Phase 3 (n = 329)	p
*Age (year)	41.09 ± 15.82	44.84 ± 15.71	47.46 ± 15.62	<0.001
†Sex (female)	202 (61.4)	202 (61.4)	202 (61.4)	1
*Low HDL-c	225 (68.4)	261 (79.3)	217 (66)	<0.001
*High WC	90 (27.4)	134 (4.7)	151 (45.9)	<0.001
*Hypertension	111 (33.7)	97 (29.5)	82 (24.9)	0.002
*High TG	139 (42.2)	140 (42.6)	141 (42.9)	0.97
*High FBS	39 (11.9)	39 (11.9)	41 (12.5)	0.85
*Smoker	27 (8.2)	30 (9.1)	27 (8.2)	0.66

Entries are mean ± sd for Age and number (%) for the rest categorical variables. *is time dependent variable. †is time independent variable.

Table 6. Genotype and allele frequencies of Apo E, Apo A1M1, Apo A1M2, Apo B, Apo AIV, Apo CIII, and SRB1 in the study population.

Polymorphisms			
Apo E Alleles	e2	e3	e4
	34 (1.3)	258 (78.4)	37 (11.2)
Apo A1M1 Genotypes	+/+	+/−	−/−
	233 (7.8)	90 (27.4)	6 (1.8)
Apo A1M2 Genotypes	+/+	+/−	−/−
	300 (91.2)	23 (7)	6 (1.8)
Apo B Genotypes	X+X+	X+X−	X−X−
	28 (8.5)	126 (38.3)	175 (53.2)
Apo AIV Genotypes	TT	GT	GG
	1 (0.3)	56 (17)	272 (82.7)
Apo CIII Genotypes	CC	CG	GG
	232 (7.5)	87 (26.4)	10 (3)
ABC A1Genotypes	GG	GA	AA
	112 (34)	171 (52)	46 (14)
SR B1 Genotypes	GG	GA	AA
	268 (81.5)	58 (17.6)	3 (0.9)

associated with an increased odds of having low HDL (OR = 2.30). The odds ratio for having low level of HDL in subjects with ((high Blood pressure and male) or (being homozygous for the minor allele of SRB1)) combination is 0.38. The first order Markov chain dependence between adjacent observations of response was estimated 2.5 indicating the strong dependence between successive observations. The AIC for this model was 100.72. Also, second order Markov chain Transition Logic Regression model with three tree logic and 8 leaves was fitted. The result of Transition Logic Regression with second order Markov chain was fairly similar to the result of the first order. According to this model, The odds ratio for having low level of HDL in subjects with ((high Blood pressure and male) or (being homozygous for the minor allele of SRB1)) combination is 0.37 and being in phase 2 or being homozygous for the minor allele of ApoCIII was associated with an increased odds of having low HDL. Also, subjects with high triglyceride that have high waist circumstance or high blood pressure have an odds ratio of 2.51 to have low level of HDL. The first order Markov chain dependence between adjacent observations of response was estimated 2.32 and the second order Markov chain dependence was 1.65. The AIC for this model was 974.96. Results of first and second-order Transition Logic Regression are shown in **Table 7**.

Table 7. Results of Transition Logic Regression model with 3 Boolean combination of 8 binary predictor variables for first and second order Markov chain dependence structure to study interaction effects of SNPs and other risk factors on having low level of HDL.

Model	Boolean combination	Odds ratio with CI 95%	AIC	$\log\psi_1$	$\log\psi_2$
First Order Transition Logic Regression	$L_1 = (\text{high TG} \wedge \text{high WC})$	2.29 (1.51, 3.48)	100.72	2.5	-
	$L_2 = (\text{phase2} \wedge (\text{Apo A1M1} \neq +/+ \vee (\text{Apo CIII} = \text{CC})))$	2.30 (1.77, 2.99)			
	$L_3 = ((\text{high BP} \wedge (\text{sex} = \text{male})) \vee \text{SRB1} = \text{AA})$	0.38 (0.25, 0.59)			
Second Order Transition Logic Regression	$L_1 = \text{high TG} \wedge (\text{high WC} \vee \text{high BP})$	2.51 (1.66, 3.78)	974.96	2.32	1.65
	$L_2 = (\text{phase2} \vee \text{ApoCIII} = \text{GG})$	2.01 (1.59, 2.54)			
	$L_3 = ((\text{high BP} \wedge (\text{Sex} = \text{male})) \vee \text{SRB1} = \text{AA})$	0.37 (0.24, 0.59)			

5. Discussion

In the first part of the paper, we extended Logic Regression and proposed a model which allowed first and second order Markov dependence in longitudinal binary data for which the marginal probability of success was modeled via a form of Logic Regression. In the second part, a simulation study was done that evaluated performance of the proposed model in different conditions. The simulation study indicated a satisfactory behavior for proposed model so that, in all condition AIC of Transition Logic Regression models were less than AIC of transition models with only main effect. Moreover, Transition Logic Regression was able to find moderate or strong interaction effects nearly in all datasets for sample sizes more than 50. In sample size 50 the quality of the estimators were poor. In this sample size, MSE of ψ_1 is not acceptable especially for strong interaction effect and high dependency. Also in this sample size, confidence intervals of estimators of β in $\beta = 0.5$ have not consisted true value of the parameter. By increasing the sample size, MSE measures of estimation for ψ_1 and β were decreased so that in other sample sizes, the performance of the method and quality of estimators are acceptable.

In the last part of the paper, proposed models were applied to the data from TLGS study and some interactions among SNPs and other covariates related to low HDL were identified. The results of first and second order Markov chain were fairly similar to each other and both of them had similar combinations. AIC of Transition Logic Regression model with second order dependency was less than the model with first order so it can be concluded that second order model is able to fit the data better than first order.

In this study, we had to work only with complete dataset because in Logic Regression methodology, missing problem has not been solved yet. It will be helpful if missing data is addressed in Logic Regression in future research.

6. Conclusion

Considering the identification of interactions in longitudinal study with binary response, Transition Logic Regression was introduced and used to find interactions influencing low HDL over time and the most important interactions were identified.

Acknowledgements

We would like to thank the staff and participants in the TLGS study for data collection. The research presented in this paper was carried out on the High Performance Computing Cluster supported by the computer science department of Institute for Research in Fundamental Sciences (IPM). This article has been extracted from PhD thesis of Parvin Sarbakhsh in Biostatistics at School of Para-medicine at Shahid Beheshti University of Medical Sciences.

Conflict of Interest

The authors have declared no conflict of interest.

References

[1] Ruczinski, I., Kooperberg, C. and Le Blanc, M. (2003) Logic Regression. *Journal of Computational and Graphical Statistics*, **12**, 475-511. http://dx.doi.org/10.1198/1061860032238

[2] Schwender, H. and Ruczinski, I. (2010) Logic Regression and Its Extensions. *Advances in Genetics*, **72**, 25-45. http://dx.doi.org/10.1016/B978-0-12-380862-2.00002-3

[3] Li, Q., Fallin, M.D., Louis, T.A., Lasseter, V.K., McGrath, J.A., Avramopoulos, D., Wolyniec, P.S., Valle, D., Liang, K.Y., Pulver, A.E. and Ruczinski, I. (2010) Detection of SNP-SNP Interactions in Trios of Parents with Schizophrenic Children. *Genetic Epidemiology*, **34**, 396-406. http://dx.doi.org/10.1002/gepi.20488

[4] Kooperberg, C. and Ruczinski, I. (2005) Identifying Interacting SNPs Using Monte Carlo Logic Regression. *Genetic Epidemiology*, **28**, 157-170. http://dx.doi.org/10.1002/gepi.20042

[5] Schwender, H. and Ickstadt, K. (2008) Identification of SNP Interactions Using Logic Regression. *Biostatistics*, **9**, 187-198. http://dx.doi.org/10.1093/biostatistics/kxm024

[6] Nunkesser, R., Bernholt, T., Schwender, H., Ickstadt, K. and Wegener, I. (2007) Detecting High-Order Interactions of Single Nucleotide Polymorphisms Using Genetic Programming. *Bioinformatics*, **23**, 3280-3288. http://dx.doi.org/10.1093/bioinformatics/btm522

[7] Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994) Analysis of Longitudinal Data. Oxford University Press, New York.

[8] Mehrabi, Y., Sarbakhsh, P., Houwing-Duistermaat, J.J., Zayeri, F. and Sadat Daneshpour, M. (2015) Assessment of SNP Interactions Affecting Total Cholesterol over Time Using Logic Mixed Model: TLGS Study. *Gene Cell Tissue*, **2**, e25572. http://dx.doi.org/10.17795/gct-25572

[9] Goncalves, M.H. and Azzalini, A. (2008) Using Markov Chains for Marginal Modelling of Binary Longitudinal Data in an Exact Likelihood Approach. *Metron-International Journal of Statistics*, **LXVI**, 157-181.

[10] Garte, S. (2001) Metabolic Susceptibility Genes as Cancer Risk Factors: Time for a Reassessment? *Cancer Epidemiology Biomarkers & Prevention*, **10**, 1233-1237.

[11] Azizi, F., Rahmani, M., Emami, H., Mirmiran, P., Hajipour, R., Madjid, M., Ghanbili, J., Ghanbarian, A., Mehrabi, J., Saadat, N., Salehi, P., Mortazavi, N., Heydarian, P., Sarbazi, N., Allahverdian, S., Saadati, N., Ainy, E. and Mœini, S. (2002) Cardiovascular Risk Factors in an Iranian Urban Population: Tehran Lipid and Glucose Study (Phase 1). *Sozial- und Präventivmedizin/Social and Preventive Medicine*, **47**, 408-426. http://dx.doi.org/10.1007/s000380200008

[12] Azizi F Fau-Ghanbarian, A., Ghanbarian A Fau-Momenan, A.A., Momenan Aa Fau-Hadaegh, F., Hadaegh F Fau-Mirmiran, P., Mirmiran P Fau-Hedayati, M., Hedayati M Fau-Mehrabi, Y., Mehrabi Y Fau-Zahedi-Asl, S. and Zahedi-Asl, S. (2009) Prevention of Non-Communicable Disease in a Population in Nutrition Transition: Tehran Lipid and Glucose Study phase II.

[13] Azizi, F., Khalili, D., Aghajani, H., Esteghamati, A., Hosseinpanah, F., Delavari, A., Larijani, B., Mirmiran, P., Mehrabi, Y., Kelishadi, R. and Hadaegh, F. (2010) Appropriate Waist Circumference Cut-Off Points among Iranian Adults: The First Report of the Iranian National Committee of Obesity. *Archives of Iranian Medicine*, **13**, 243-244.

[14] NCEP (2001) Executive Summary of the Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *JAMA*, **285**, 2486-2497. http://dx.doi.org/10.1001/jama.285.19.2486

[15] NCEP (2001) Executive Summary of the Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *JAMA*, **285**, 2486-2497. http://dx.doi.org/10.1001/jama.285.19.2486

[16] Daneshpour, M.S., Faam, B., Hedayati, M., Eshraghi, P. and Azizi, F. (2011) ApoB (XbaI) Polymorphism and Lipid Variation in Teharnian Population. *European Journal of Lipid Science and Technology*, **113**, 436-440. http://dx.doi.org/10.1002/ejlt.201000346

[17] Brown, C.M., Rea, T.J., Hamon, S.C., Hixson, J.E., Boerwinkle, E., Clark, A.G. and Sing, C.F. (2006) The Contribution of Individual and Pairwise Combinations of SNPs in the APOA1 and APOC3 Genes to Interindividual HDL-C Variability. *Journal of Molecular Medicine*, **84**, 561-572.

[18] Daneshpour, M.S., Hedayati, M., Eshraghi, P. and Azizi, F. (2010) Association of Apo E Gene Polymorphism with HDL Level in Tehranian Population. *European Journal of Lipid Science and Technology*, **112**, 810-816. http://dx.doi.org/10.1002/ejlt.200900207

[19] McCarthy, J.J., Lehner, T., Reeves, C., Moliterno, D.J., Newby, L.K., Rogers, W.J. and Topol, E.J. (2003) Association of Genetic Variants in the HDL Receptor, SR-B1, with Abnormal Lipids in Women with Coronary Artery Disease. *Journal of Medical Genetics*, **40**, 453-458. http://dx.doi.org/10.1136/jmg.40.6.453

[20] Frikke-Schmidt, R. (2000) Context-Dependent and Invariant Associations between APOE Genotype and Levels of Lipoproteins and Risk of Ischemic Heart Disease: A Review. *Scandinavian Journal of Clinical and Laboratory Investigation*, **233**, 3-25.