行政院國家科學委員會專題研究計畫 成果報告

戒菸期的傾向:一個馬可夫的方法
研究成果報告(精簡版)

計畫類別：個別型
計畫編號：NSC 98-2118-M-009-005-
執行期間：98年08月01日至99年07月31日
執行單位：國立交通大學統計學研究所

計畫主持人：彭南夫

計畫參與人員：碩士班研究生-兼任助理人員：沈彥廷
碩士班研究生-兼任助理人員：洪鏡婷

處理方式：本計畫可公開查詢

中華民國 99年10月28日

Trends in Smoking Cessation: A Markov Approach
戒菸期的傾向 ： 一個馬可夫的方法

中文摘要
在戒菸實驗中，我們常會觀察到參與者有多重離散型的階段。過去已有一些關於此的長期資料分析，我們也可假設這些資料具有馬可夫的性質。參與實驗者常會有驅向某一階段的傾向，我們用對數轉換傾向參數模型，提出別於二元的新的模型與方法，做出估計與檢定的問題。這些新的模型與方法將用於戒菸實驗的數據。我們也會做模擬實驗。

關鍵字 ： 馬可夫鍊 ， 長期資料 ， 戒煙

Abstract
Intervention trials such as studies on smoking cessation may observe multiple, discrete outcomes over time. Participant observations may alternate states over the course of a study. Approaches exist which are commonly used to analyze binary, longitudinal data in the context of independent variables. However, the sequence of observations may be assumed to follow a Markov chain with stationary transition probabilities when observations are made at fixed time points. Participants favoring the transition to one particular state over the others would evidence a trend in the observations. Using a log-transformed trend parameter, the determinants of a trend in a binary, longitudinal study may be evaluated by maximizing the likelihood function. New methodology on extension to discrete time Markov chain model and continuous time Markov chain model is studied here to test for the presence and determinants of a trend in multiple state, rather than binary, longitudinal observations. Practical application of the proposed method is made to data available from an intervention study on smoking cessation. Simulation studies will also be taken.

Keywords: Markov chains, longitudinal data, smoking cessation

I. Introduction

Discrete outcomes are repeatedly measured in many areas of research. When the observation is binary, study participants may alternate between two classes (or states) over the course of a study. Subjects may tend to favor the transition to a particular state depending on known or unknown factors. For instance, the goal of a smoking cessation program may be to influence participants’ decisions in favor of smoking abstinence, as opposed to relapse. Study investigators may hypothesize a trend toward the abstinence state depending on the intervention group or covariates of interest.

Other example of discrete, longitudinal data may include infection status, psychological state, or drug therapy compliance (Liu and others, 1995; Dascalakis and others, 2002; Solomon and others, 2005).

Traditional analysis, such as logistic regression, are inappropriate when repeated measurements are made on the same subject due to an inherent correlation in the measurements. Generalized linear mixed models are commonly used to model discrete, longitudinal outcomes using random effects (Molenberghs and Verbeke, 2005). However, parameter estimates under this model are subject-specific (Molenberghs and Verbeke, 2005). A methodology that focuses on population average parameters is the generalized estimating equations (GEE) approach (Liang and Zeger, 1986; Zeger and Liang, 1986; Hu and others, 1998; Hardin and Hilbe, 2003). Discrete time Markov models may also be used to analyze binary longitudinal data (Li and Chan, 2006).

Generalized estimating equations (GEE) are commonly used to analyze binary, longitudinal data in the context of independent variables (Liang and Zeger, 1986; Zeger and Liang, 1986). Under the GEE model, a trend may be defined as a change in the log-odds of an event per unit increase in time. Statistically significant covariate-time interactions imply that a trend in the log-odds of an event is covariate dependent.

Previous work describing binary, longitudinal data through Markov models has focused on logistic regression methods to model the transition probabilities and transition model that do not allow for the inclusion of covariates (Rieger, 1968; Muenz and Rubinstein, 1985). Corcoran and others (2001) suggest an exact trend test for correlated, binary data. However, this test may be computationally infeasible in the presence of continuous covariates or with the inclusion os multiple covariates (Corcoran and others, 2001). A continuous time, binary Markov model has previously described that models the transition probabilities as exponential functions of covariates (Jones and others, 2006). However, trend in the response variables is not evaluated. This project extends the methodologies presented by Regier (1968) to identify the presence and determinants of trend in multiple, rather then binary, longitudinal data for multiple subjects (or groups). The methodology developed utilizes a stationary, multi-state Markov chain for N subjects in the context of a log-transformed trend parameter described as a linear function of covariates. The likelihood function is described, maximum likelihood estimates of unknown parameters are searched, and a likelihood ratio test is used to test hypothesis about trend.

II. Methods

- **Assumptions**

The outcome of interest is assumed to be a repeatedly measured multiple-state variable. Pair-wise observations within a subject may be correlated, but observations across subjects are assumed to be independent. Transition probabilities are assumed to depend only on the current state of each subject, and otherwise do not depend on previously recorded observations. The proposed Markov model assumes that a log-transformed trend parameter is a linear function of covariates. For discrete time Markov model, the length of time between observations is one time unit. For continuous time Markov model, the observation times are $t_{1i},\cdots,t_{Mi}$ for subject *i*.

- **Likelihood Function**

1. **Discrete Time Markov Chain Model**

An extension of the model of Regier (1968) defines a transition probability matrix that describes the multi-state transition probabilities for a single subject with respect to a trend parameter (θ). When *N* subjects are observed, let θ_i describe the trend for subject *i* $(i=1,\cdots,N)$. Let $p=(p_1,\cdots,p_m)$, the transition probability matrix may then be written as

$$T_{i1}=$$

$$\begin{bmatrix} \frac{\theta_i p_{\tau_{11}}}{\theta_i p_{\tau_{11}}+p_{\tau_{12}}+\cdots+p_{\tau_{1m}}} & \frac{p_{\tau_{12}}}{\theta_i p_{\tau_{11}}+p_{\tau_{12}}+\cdots+p_{\tau_{1m}}} & \cdots & \frac{p_{\tau_{1m}}}{\theta_i p_{\tau_{11}}+p_{\tau_{12}}+\cdots+p_{\tau_{1m}}} \\ \frac{\theta_i p_{\tau_{21}}}{\theta_i p_{\tau_{21}}+p_{\tau_{22}}+\cdots+p_{\tau_{2m}}} & \frac{p_{\tau_{22}}}{\theta_i p_{\tau_{21}}+p_{\tau_{22}}+\cdots+p_{\tau_{2m}}} & \cdots & \frac{p_{\tau_{2m}}}{\theta_i p_{\tau_{21}}+p_{\tau_{22}}+\cdots+p_{\tau_{2m}}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\theta_i p_{\tau_{m1}}}{\theta_i p_{\tau_{m1}}+p_{\tau_{m2}}+\cdots+p_{\tau_{mm}}} & \frac{p_{\tau_{m2}}}{\theta_i p_{\tau_{m1}}+p_{\tau_{m2}}+\cdots+p_{\tau_{mm}}} & \cdots & \frac{p_{\tau_{mm}}}{\theta_i p_{\tau_{m1}}+p_{\tau_{m2}}+\cdots+p_{\tau_{mm}}} \end{bmatrix}$$

where $(\tau_{i1},\cdots,\tau_{im})$ is a known permutation of $(1,\cdots,m)$ depending on cases. In this matrix θ_i describes the change in the transition probabilities for the *i*th individual, and *p* describes the population transition probabilities in the absence of a trend. Let $\ln(\theta_i)$ be the link function that maps the transition trend to the linear function

$$\ln(\theta_i)=X_i^{'}\beta$$

where $X_i^{'}$ is the vector of covariates for subjects *i* and β is the vector of population parameters. Each element of T_{i1} describes the one-step transition probabilities in a sequence of observations. Let *R* be the matrix of observed sequences for *N* subjects

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1M} \\ r_{21} & r_{22} & \cdots & r_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N1} & r_{N2} & \cdots & r_{NM} \end{bmatrix}$$

where r_{ij} stands for the state of subject *i* at time *j*. The likelihood function appears to be

$$L(\beta, p \,|\, X, R) = \prod_{i=1}^{N} \prod_{j=1}^{m} \prod_{k=1}^{m} (T_{i1}[j,k])^{n_{ijk}}$$

where n_{ijk} is the number of transitions between state *j* and state *k* for subject *i*.

This likelihood function must be modified in the presence of missing data to include n-step transition probabilities (Howard and Karlin 1998). Missing data may occur at the beginning , middle or end of each sequence of observations. Data missing at the beginning or end of each sequence may be ignored with respect to the likelihood function as such missingness does not contribute any information about the transitions. However, information is contained in the observations when missing data are present in the middle of a sequence, between two observed data points. When one data point is missing between two observations, the 2-step transition probability matrix is $T_{i2} = T_{i1}{}^{2}$. A similar fashion can be extended to *n*-step missing data.

2. Continuous Time Markov Chain Model

We here assume that the infinitesimal matrix

$$Q = \begin{bmatrix} \Delta & \theta_i p_{\tau_{12}} & \cdots & \theta_i p_{\tau_{1m}} \\ \theta_i p_{\tau_{21}} & \Delta & \cdots & p_{\tau_{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_i p_{\tau_{m1}} & p_{\tau_{m2}} & \cdots & \Delta \end{bmatrix}$$

where for simplicity the Δ in *Q* is a symbol that the sum of each row of *Q* is zero. In order to obtain the maximum likelihood estimates of the parameters, it is essential to find the transition probability matrix $T_{i1}(t_{(k+1)i} - t_{ki})$ between two observation times t_{ki} and $t_{(k+1)i}$ in advance. Such a problem can be solved numerically by the following identity,

Proposition (Ross, 1996)

$$T_{i1}(t_{(k+1)i} - t_{ki}) = \lim_{m \to \infty} (I + \frac{Q(t_{(k+1)i} - t_{ki})}{2^m})^{2^m} .$$

For example, *k* = 20 matrix multiplications can reach high accuracy. Once the transition probability matrix of continuous time Markov chain is obtained, the rest part of the task required is similar to that of discrete time Markov model. That is,

- **Maximum Likelihood Estimation**

Maximizing the logarithm of the likelihood function with respect to the unknown parameters (β and p) is computationally difficult, especially in the presence of missing data. Therefore, the Nelder-Mead simplex algorithm, which does not require differentiation, is used to find maximum likelihood estimates (Nelder and Mead, 1965). SAS is used to implement this optimization using the NLPNMS call function. Parameter constrains are required for p_i to prevent division by zero; therefore, $0.00001 \le p_i \le 0.99999$.

- **Simulation Studies**

This study focuses on testing for statistically significant associations between trend and a binary covariate (e.g. placebo versus experimental treatment groups). The log-transformed trend parameter is modeled as

$$\ln(\theta_i) = \beta_0 + \beta_1 \ X_i \quad ; i = 1, \cdots, N$$

under the proposed Markov model. The null hypothesis of interest is

$$H_0 : \beta_1 = 0 .$$

Note that β_1 represents the change in the $\ln(\theta)$ for *X* = 1 compared to *X* = 0.

III. Application of Models to Investigate Smoking Relapse Trend in the Smoking Treatment Study

Unpublished data is available from an intervention study on smoking cessation. Dr. David Wetter, The University of Texas M.D. Anderson Cancer Center in Houston, TX, initiated a study entitled “Smoking Treatment on Palmtops” (STOP) to evaluate the effectiveness of computer delivered therapy on smoking cessation. Two study sites were established to enroll participants (Houston, TX, and Seattle, WA) from September 2000 through June 2002. Participants (N = 303) were randomly assigned to a control group or experimental treatment group. The control group received standard therapy including nicotine patches and self-help materials. The experimental group received a computer delivered treatment in addition to standard therapy.

IV. Discussion

For both models, power increases as th number of observations per subject increases. Parameter estimates are not explicitly comparable. Under the proposed Markov model, several limitations exist. First, transition probabilities are assumed to be stationary, and the proposed method is inappropriate when the outcome variables follow a non-stationary process. The Nelder-Mead simplex algorithm is used to maximize the likelihood functions. This algorithm may fail to converge, or may not converge to global maximum. However, most of the calculation converge in this simulation.

References

Corcoran,C., Ryan, L., Sanchaudhuri, P. Mehta, Patel, N.(2001) An exact trend test for correlated binary data. Biometrics, 57(3):941-948

Daskalakis, C, Laird, N. Murthy, J. (2002) Regression analysis of multiple source longitudinal outcomes: A stirling county depression study, American journal of Epidemiology, 155(1) 88-94

Hardin,J. and Hilbe, J. (2003)Generalized Estimating Equations, Chapman and Hall

Hu,F., Goldger,J. Hedeker, D., Flay,B.(1998) Comparison of population averaged and subject specific approaches for analyzing repeated binary outcomes. American journal of Epidemiology 147, (7), 694-703

Jones, R.,Xu,S.and Grunwald, G.. (2006) Continuous time Markov models for binary longitudinal data, Biometrical journal, 48(3), 411-419

Li,Y. and Chan,W.(2006) Analysis of longitudinal multinomial outcome data, Biometrical journal, 48(2), 319-326

Liang K. and Zeger, S. (1986) Longitudinal data analysis using generalized linear models. Biometrika, 73, 13-22

Liu, T. Soong, S. Alvarez, R. Butterworth, C. (1995) A longitudinal analysis of human papillomavirus 16 infection, nutritional status, and cervical dysplasia progression. Cancer Epidemiology, Biomarkers and Prevention 4(4), 373-380

Molenberghs, G. and Verbeke, G. (2005) Models for discrete longitudinal data. Springer, NY

Muenz, L. and Rubinstein,.L.(`1995) Markov models for covariate dependenceof binary sequences. Biometrics 41(1), 91-101

Regier, M. (1968) Two state Markov model for behavioral change, Journal of American Statistical Association. 63, 993-999

Ross, S. (1996) Stochastic Processes, 2nd ed., Wiley

Solomon, D. Avron, J. Katz ,J. Finklestein, J. (2005) Compliance with osteoporosis medications. Archives of Internal Medicine, 165 (20) 2414-2419

Zeger, S. and Liang, K. (1986) Longitudinal data analysis for discrete and continuous outcomes, Biometrics, 42, 121-130.

無衍生研發成果推廣資料

98 年度專題研究計畫研究成果彙整表

計畫主持人：彭南夫	計畫編號：98-2118-M-009-005-
計畫名稱：戒菸期的傾向:一個馬可夫的方法	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無

🟥	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：□已發表 ■未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：（以 100 字為限）

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）

本研究對長期實驗性質的研究，提供一個統計分析的方法。這是一個利用馬可夫模型的方法，解決研究上的問題。