國 立 交 通 大 學

統計學研究所

碩 士 論 文

轉移機率含因子的分析方法

Analysis of Transition Probability with Covariates

研 究 生 ：艾雪芳

指導教授 ：彭南夫 博士

中 華 民 國 九 十 六 年 六 月

轉移機率含因子的分析方法

Analysis of Transition Probability with Covariates

研 究 生 ：艾雪芳　　Student: Hsueh-Fang Ai
指導教授 ：彭南夫 博士　　Advisor: Dr. Nan-Fu Peng

國 立 交 通 大 學

統計學研究所

碩 士 論 文

A Thesis

Submitted to institute of Statistics

College of Science

National Chiao Tung University

in Partial fufillment of the Requirements

for the Degree of

Master

in

Statistics

June 2007

Hsinchu, Taiwan, Republic of Chian

中 華 國 國 九 十 六 年 六 月

Analysis of transition probability with covariates

Student: Hsueh-Fang Ai Advisor: Dr. Nan-Fu Peng

Institute of Statistic

National Chiao Tung University

Abstract

In Huang et al [6] the single rate research is provided by a lot of statistical methods. Now we want to develop a new method to analyze the transition data with risk factors. With the conditional Markov model we can solve the dependent data question.

Use this method we need a large “enough” sample in each cell of transition probability. So we recode the continuous factors and use some tests to find more influential factors.

Under the model, bootstrap method can help us to construct a confidence interval for each parameter. Thus we still can test the parameter if it is significant.

The results include all the transition probability of the process of the disease. Then we can compare the rates among the factor. In this report we use ARM (age-related maculopathy) data to demonstrate the method and provide the previous research in analysis of ARM data.

Keyword: ARM, transition probability and conditional Markov Chain

轉移機率含因子的分析方法

研究生： 艾雪芳　　指導教授： 彭南夫 博士

國立交通大學統計學研究所

摘要

在過去的文獻[6]中提供了很多針對單一比率(發生率、惡化率…)的統計方法，而今我們想要發展的分析方法是針對含有風險因子的轉移資料(transition data with risk factors). 在條件馬可夫模型下我們可以解決資料不獨立的問題。

使用這個方法，我們需要足夠大的樣本所以我們把一些連續型的變數重新編碼，且用一些檢定找出比較有影響的因子，來建立統計模型。

在模型下，我們使用拔靴法(bootstrap method)來估計每個參數的信賴區間，據此來檢定各參數是否顯著。

最後我們可以看到這個疾病所有歷程上的轉移機率。所以我們可以做不同的比率(發生率、惡化率…)在不同的風險因子之間的比較。在本研究中，我們使用退化性視網膜黃斑部病變(ARM)的資料來展示這個方法，且提供之前對 ARM 資料分析的回顧。

關鍵字：退化性視網膜黃斑部病變、轉移機率、條件馬可夫鏈

誌謝

「…要感謝的人太多那就謝天吧。」(摘錄自陳之藩《謝天》)也許一連串的好運就從交大前面的土地公公開始，讓我考上的交大統研所，開始了我最想唸的領域。不知不覺兩年就過去了，我很開心有走進了統計的世界。要學的東西很多，但每學到一點又會多想知道一點，我還是第一次對於讀書有這樣的渴望。而今又幸運的考上博班，可以在我最喜歡的學門中探索更深入的學問，真是一件令人振奮的事。

這兩年先要感謝老師們的栽培，尤其是彭南夫老師對我論文的指導及黃冠華老師給寶貴的建議。另外何其有幸能與大家成為同學，分享學習及生活上的點點滴滴。感謝大家平日忍受我焦慮的啐啐唸，生日時送我小時候想看的三角系列小說。時間過得真快，從完全不懂數統到開始唸統計，從超級路痴到可以幫人指路，羽球也進步到會殺球，好像都有點長進，除了數十年如一日的牌技外。

謝謝火哥、穗碧學姊、小米、素梅、大哥大嫂、映伶、阿淳和雅莉，因為有你們，研究才會這麼有趣。尤其是小米，你真是一個可以好好討論問題的同學，不管是隨機、倖存還是實驗設計。最後謝謝從小到大的朋友，謝謝你們平日聽我倒垃圾。常常陪我看電影，在沒有公車時跑來載我回家的 Amy 小姐；另外還有常常沒事來吵我，要我載他回家的小華。謹以本論文獻給我的師長、朋友及同學，並獻上我最誠摯的謝意。

艾雪芳 丁亥夏至

Content

Tables and Figures Content

1 Introduction

1.1 Purpose of Research

The study we present is a new analysis of transition data with factors by using Markov chain. Select the factors to build a statistical model and will pay more attention. In the cohort study we will follow the same participants in each time echo however the information they provide will dependent. Today we add a Markov property assumption to solve difficulty in the analysis of dependent data. Conditional on covariates the process of each person will be a Markov chain to solve difficulty in the analysis of dependent data.

The new strait in the analysis is that we need an “enough” sample size. In this example with many factors then the size of subgroup is small not mention to the continuous factors.

Our method is to break the continuous data to categorical data simplified the level of factor and use homogeneity and stationarity tests to pick up the influential factors.

1.2 Background

Vision composition

The visional system consists of retina of eyes which connected optic nerve to the visional center of brain. The maculopathy on the retina is against to the pupil with 5.5 mm larger than the pupil. If we look straightly then the maculopathy can controls 20 degree of viewpoint. Highly sensitive visional cells are located at the maculopathy although its area occupied the all retina is just 2%. The visional center uses more then one half of cells to analyze the information by it received.

The vision consists of retina and the maculopathy is not only the geographic center but also the center of visional center. Once the maculopathy has pathologies then the vision is

effected by it even loosed one's sight.

1.3 Literature review

The previous researches in the Beaver Dam Eye Study [7-17] provide some statistical method but most of them are focus on the single events rate such prevalence rate or procession rate et al. Our method in this paper not only solve the dependent data but also provide a global view to understand how a factor/factors act on the ARM.

These risk factors were chosen because of a strong relation with age-related maculopathy in previous studies. In the Beaver Dam Eye Study, smoking was related to the prevalence of age-related maculopathy [12], heavy drinking and hypertension were associated with exudative macular degeneration, a lesion that defined late age-related maculopathy [14, 16], and serum cholesterol was inversely associated with age-related maculopathy [10]. Vitamin use was found to be associated with the incidence of age-related maculopathy in a clinical trial [1]. Definitions of these confounding variables have been described in detail elsewhere [2, 12, 14 and 19]. In brief, a subject was classified as a current smoker if he/she had smoked more than 100 cigarettes in his/her lifetime and had not stopped smoking; as a former smoker if he/she had smoked more than this number but had not smoked within the last year prior to the examination; and as a nonsmoker if he/she had smoked fewer than 100 cigarettes in his/her lifetime. A current heavy drinker was defined as a person consuming four or more servings of alcoholic beverages daily, a former heavy drinker had consumed four or more servings daily in the past but not within the last year, and a non-heavy drinker had never consumed four or more servings daily on a regular basis. A person was classified as a current vitamin user if he/she had taken at least one vitamin per week in the month prior to the examination; as a past vitamin user if he/she had ever regularly taken vitamins at least once a week but not within the last month; and as never using vitamins if she/she never took vitamins regularly. Hypertension was defined as

a systolic blood pressure of 160 mmHg and/or a diastolic blood pressure of 95 mmHg and/or a history of hypertension using antihypertensive medication at the time of the examination. We adjusted for these potential confounding variables in each model. Measurements of risk factors were taken at each examination; however, multivitamin use and cholesterol level were not available at the 10-year follow-up. In the following analysis, we use the 5-year multivitamin use and cholesterol level as the 10-year measurements.

The possible reasons for nonparticipation include death, moving out of the area, and refusal [9, 11 and 13] .Comparisons between participants and non-participants at all three examinations have been presented elsewhere [9, 11 and 13].

1.4 Procedures

Procedures for obtaining and evaluating photographs of participants' eyes have been described elsewhere [9, 17]. At each examination, 30° color stereoscopic fundus photographs were taken of both of each participant's eyes. Preliminary and detailed grading was then carried out on the fundus photographs to determine the presence and severity of specific lesions associated with age-related maculopathy, including largest drusen size, most severe drusen type (in order of increasing severity: hard distinct drusen, soft distinct drusen, soft indistinct drusen, and reticular drusen), increased retinal pigment, retinal pigment epithelial depigmentation, exudative macular degeneration (retinal pigment epithelial detachment, subretinal hemorrhage, subretinal fibrosis), and geographic atrophy.

In this reporter, we adopt 6-level scale. Experienced graders used the photographs to evaluate the severity of lesions of ARM, which were graded on a 6-level scale, such as 10 20 … 60 [14].

The scale will be re-classified to three levels (the detail is the following definition), in order to increase severity: level 0 = disease free, level 1 = early ARM and level 2 = late ARM. Results presented here use each individual's ARM level in the worse eye. [6]

Define 1.1 the states of ARM

Level 0: disease free if 6-level=10
Level 1: early ARM if 6-level=20/30/40
Level 2: late ARM if 6-level=50/60

2 Methodology

2.1 Statistical model

We want to model the transition probability with covariates. The probability form state *i* to state *j* is

$$P_{ij}^{t} = \exp\left(\beta_0 + \sum_{m=1}^{p} \beta_m^t x_m^t + \sum_{m \neq n} \gamma_{m,n}^t x_m^t x_n^t \right)_{ij}$$

where *p* is the number of covariates and the *t* is the time echo.

x_m : covariate (risk factor)
β_0 : minus log transition prob. of based line
β_m : decrese the log transition prob. for specified risk factor *m*
γ_{mn} : decrese the log transition prob. for risk factor *m* and *n*

The procedure is as following.

1. Pick up the complete data in all time echoes.
2. Test each covariate if it is significant.
3. Pick up the complete data again focus on the significant covariates.
4. Estimate the mle and the confidence intervals.

2.2 Select the influential factors

The Markov chain needs an enough sample size and uniform data to construct the model so we hope just select influential factors which can affect the over all transition. The first thing we have to do is to scan all the factors and put the influential ones in our model. The follow tests [3, 18] will help us select the influential factors.

Test for stationarity of the Transition probability Matrix

a. The approximant distribution of likelihood ratio

If $L(\mathbf{P})$ is the likelihood function, $l(\mathbf{P}) = \ln(L(\mathbf{P}))$

Λ is the likelihood ratio then

$$-2\ln(\Lambda) \sim \chi^2_{d.f.}$$

If the transition probability with *m* states then

$$-2\ln(\Lambda) = 2\left[l(\hat{\mathbf{P}}) - l(\mathbf{P}^0)\right] = 2\sum_{i=1}^{m}\sum_{j=1}^{m} n_{ij} \ln\left(\frac{n_{ij}}{n_{i.}P_{ij}^0}\right) \sim \chi^2$$

The degree of freedom is *m* (*m-1*)

b. Test for stationarity

At time *t* the transition probability is $P_{ij}^t = P\left[X(t+1) = j \mid X(t) = i\right]$

	1	2	⋯	M
1	n_{i1}^1	n_{i2}^1	⋯	n_{im}^1
2	n_{i1}^2	n_{i2}^2		n_{im}^2
⋮	⋮	⋮		⋮
T	n_{i1}^T	n_{i2}^T	⋯	n_{im}^T

$$H_0 : P_{ij}^t = P_{ij} \ (t = 1,2 \cdots T)$$

$$-2\ln(\Lambda) = 2\left[L(\hat{\mathbf{p}}^t) - L(\hat{\mathbf{p}})\right] = 2\sum_{t=1}^{T}\sum_{i=1}^{m}\sum_{j=1}^{m} n_{ij}^t \ln\frac{n_{ij}^t}{n_i^{t-1}P_{ij}} \sim \chi^2_{(T-1)m^2}$$

Test for homogeneity across several Markov Chain

If there are *S* samples of Markov Chain, each transition probability is

$P_{ij}^l \ l \in \{1,\cdots,S\}$ and n_{ij}^l : transition count and $n_{i.}^l = \sum_{j=1}^{m} n_{ij}^l$

Each mle of different sample is $\hat{P}_{ij}^l = \frac{n_{ij}^l}{n_{i.}^l}$ and the mle of pooled transition probability

$$\hat{P}_{ij} = \frac{\sum_{l=1}^{S} n_{ij}^l}{\sum_{l=1}^{S} n_{i.}^l}$$

$$H_0 : P_{ij}^1 = P_{ij}^2 \cdots = P_{ij}^S$$

$$\sum_{l=1}^{S}\sum_{i}\sum_{j}\frac{\left(n_{ij}^l - n_{i.}^l\hat{P}_{ij}\right)^2}{n_{i.}^l\hat{P}_{ij}} \sim \chi^2_{(S-1)m(m-1)}$$

2.3 Estimate the parameters

The next we want to find the mle accord to the model.

Now we have *N* individual in *t* time, x_k^t is a special factor in k_{th} individual dependant on *t*, but its parameter β_{ij} does not dependant on *t*.

We will democrat I=J=3 for some state *i* will transient to next state.

$$\mathbf{P^t} = \left[p_{ij}{}^t\right] = \begin{bmatrix} 1-\sum_{i=1} p_{ij}{}^t & b_{12}\exp\left(-\beta_{12}x^t\right) & b_{13}\exp\left(-\beta_{13}x^t\right) \\ b_{21}\exp\left(-\beta_{21}x^t\right) & 1-\sum_{i=2} p_{ij}{}^t & b_{23}\exp\left(-\beta_{23}x^t\right) \\ b_{31}\exp\left(-\beta_{31}x^t\right) & b_{32}\exp\left(-\beta_{32}x^t\right) & 1-\sum_{i=3} p_{ij}{}^t \end{bmatrix}$$

$$\mathrm{p}_{ij,k}^t = b_{ij}\exp\left(-\beta_{ij}x_k^t\right) \qquad \mathrm{p}_{ii,k}^t = 1-\sum_{i\neq j} b_{ij}\exp\left(-\beta_{ij}x_k^t\right)$$

The partial likelihood function is

$$L(\boldsymbol{\beta},\mathbf{b})_{i=1} = \prod_{t=1}^{3}\prod_{k=1}^{N}\mathbf{I}^t(1,2)b_{12}\prod_{t=1}^{3}\prod_{k=1}^{N}\mathbf{I}^t(1,3)b_{13}\exp\left(-\beta_{12}\sum_{t=1}^{3}\sum_{k=1}^{N}\mathbf{I}^t(1,2)x_k^t - \beta_{13}\sum_{t=1}^{3}\sum_{k=1}^{N}\mathbf{I}^t(1,3)x_k^t\right)$$
$$\left(\prod_{t=1}^{3}\prod_{k=1}^{N}\left(1-b_{12}\exp\left(-\beta_{12}\mathbf{I}^t(1,1)x_k^t\right)-b_{13}\exp\left(-\beta_{13}\mathbf{I}^t(1,1)x_k^t\right)\right)\right)$$

The partial log-likelihood function is

$$l(\boldsymbol{\beta},\mathbf{b})_{i=1} = \sum_{t=1}^{3}\sum_{k=1}^{N}\mathbf{I}^t(1,2)\ln(b_{12}) + \sum_{t=1}^{3}\sum_{k=1}^{N}\mathbf{I}^t(1,2)\ln(b_{13}) - \beta_{12}\sum_{t=1}^{3}\sum_{k=1}^{N}\mathbf{I}^t(1,2)x_k^t - \beta_{13}\sum_{t=1}^{3}\sum_{k=1}^{N}\mathbf{I}^t(1,3)x_k^t$$
$$+\sum_{t=1}^{3}\sum_{k=1}^{N}\ln\left(1-b_{12}\exp\left(-\beta_{12}\mathbf{I}^t(1,1)x_k^t\right)-b_{13}\exp\left(-\beta_{13}\mathbf{I}^t(1,1)x_k^t\right)\right)$$

Define:

$$X_{ij} = \sum_{t=1}^{3}\sum_{k=1}^{N}\mathbf{I}^t(i,j)x_k^t\,, \quad n_{ij} = \sum_{t=1}^{3}\sum_{k=1}^{N}\mathbf{I}^t(i,j)$$

We will solve the mle by differentiation.

Let $\dfrac{\partial\, l(\boldsymbol{\beta},\mathbf{b})_{i=1}}{\partial\, \beta_{12}} = 0$.

$$0 = -X_{12} + \sum_{t=1}^{3}\sum_{k=1}^{N} \frac{\mathbf{I}^t(1,1)x_k^t \hat{b}_{12}\exp\left(-\hat{\beta}_{12}\mathbf{I}^t(1,1)x_k^t\right)}{1-\hat{b}_{12}\exp\left(-\hat{\beta}_{12}\mathbf{I}^t(1,1)x_k^t\right)-\hat{b}_{13}\exp\left(-\hat{\beta}_{13}\mathbf{I}^t(1,1)x_k^t\right)}$$

$$X_{12} = \sum_{t=1}^{3}\sum_{k=1}^{N} \frac{\mathbf{I}^t(1,1)x_k^t \hat{b}_{12}\exp\left(-\hat{\beta}_{12}\mathbf{I}^t(1,1)x_k^t\right)}{1-\hat{b}_{12}\exp\left(-\hat{\beta}_{12}\mathbf{I}^t(1,1)x_k^t\right)-b_{13}\exp\left(-\hat{\beta}_{13}\mathbf{I}^t(1,1)x_k^t\right)}$$

$$= \sum_{\mathbf{I}(1,1),x=1} \frac{\hat{b}_{12}\exp\left(-\hat{\beta}_{12}\right)}{1-\hat{b}_{12}\exp\left(-\hat{\beta}_{12}\right)-\hat{b}_{13}\exp\left(-\hat{\beta}_{13}\right)}$$

Define:

$$\hat{P}_{ij} = \hat{b}_{ij}\exp\left(-\hat{\beta}_{ij}\right)$$

$n_{ij,x}$: The count from state i to state j conditional on some x

$$X_{13} = \sum_{\mathbf{I}(1,1),x=1} \frac{\hat{b}_{13}\exp\left(-\hat{\beta}_{13}\right)}{1-\hat{b}_{12}\exp\left(-\hat{\beta}_{12}\right)-\hat{b}_{13}\exp\left(-\hat{\beta}_{13}\right)} = \sum_{\mathbf{I}(1,1),x=1} \frac{\hat{p}_{13}}{1-\hat{p}_{12}-\hat{p}_{13}} = n_{11,x=1}\frac{\hat{p}_{13}}{1-\hat{p}_{12}-\hat{p}_{13}}$$

$$\Rightarrow \frac{X_{13}}{n_{11,x=1}} = \frac{\hat{p}_{13}}{1-\hat{p}_{12}-\hat{p}_{13}}$$

$$\Rightarrow \frac{X_{13}}{n_{11,x=1}}\left(1-\hat{p}_{12}\right) = p_{13}\left(1+\frac{X_{13}}{n_{11,x=1}}\right)$$

$$\hat{p}_{13} = \frac{X_{13}\left(1-\hat{p}_{12}\right)}{\left(n_{11,x=1}+X_{13}\right)}$$

$$\Rightarrow 1-\hat{p}_{13} = \frac{n_{11,x=1}+X_{13}-X_{13}\left(1-\hat{p}_{12}\right)}{\left(n_{11,x=1}+X_{13}\right)} = \frac{n_{11,x=1}-X_{13}\hat{p}_{12}}{\left(n_{11,x=1}+X_{13}\right)}$$

$$\hat{p}_{12} = \frac{X_{12}\left(1-\hat{p}_{13}\right)}{n_{11,x=1}+X_{12}}$$

$$\Rightarrow \hat{p}_{12}\left(n_{11,x=1}+X_{12}\right) = X_{12}\left(1-\hat{p}_{13}\right)$$

$$\Rightarrow \hat{p}_{12}\left(n_{11,x=1}+X_{12}\right) = X_{12}\frac{n_{11,x=1}-X_{13}\hat{p}_{12}}{\left(n_{11,x=1}+X_{13}\right)}$$

$$\Rightarrow \hat{p}_{12}\left(n_{11,x=1}+X_{12}-\frac{X_{12}X_{13}}{\left(n_{11,x=1}+X_{13}\right)}\right) = \frac{X_{12}n_{11,x=1}}{\left(n_{11,x=1}+X_{13}\right)}$$

$$\Rightarrow \hat{p}_{12}\left(\left(n_{11,x=1}+X_{12}\right)\left(n_{11,x=1}+X_{13}\right)-X_{12}X_{13}\right) = n_{11,x=1}X_{12}$$

$$\Rightarrow \hat{p}_{12} = \frac{n_{11,x=1} X_{12}}{n_{11,x=1}\left(n_{11,x=1} + X_{12} + X_{13}\right)} = \frac{X_{12}}{n_{11,x=1} + X_{12} + X_{13}} = \frac{X_{12}}{X_{11} + X_{12} + X_{13}}$$

$$\hat{p}_{12} = \frac{X_{12}}{X_{11} + X_{12} + X_{13}}$$

then as the same from $\hat{p}_{13} = \dfrac{X_{13}}{X_{11} + X_{12} + X_{13}}$

The next

$$\frac{\partial\, l(\boldsymbol{\beta}, \mathbf{b})_{i=1}}{\partial\, b_{12}} = 0$$

$$0 = \frac{n_{12}}{\hat{b}_{12}} + \sum_{t=1}^{3}\sum_{k=1}^{N} \frac{-\exp\left(-\hat{\beta}_{12}\mathbf{I}^t(1,1)x_k^t\right)}{1-\hat{b}_{12}\exp\left(-\hat{\beta}_{12}\mathbf{I}^t(1,1)x_k^t\right)-\hat{b}_{13}\exp\left(-\hat{\beta}_{13}\mathbf{I}^t(1,1)x_k^t\right)}$$

$$\Rightarrow \frac{n_{12}}{\hat{b}_{12}} = \sum_{\mathbf{I}(1,1),x=0}\left(\frac{1}{1-\hat{b}_{12}-\hat{b}_{13}}\right) + \sum_{\mathbf{I}(1,1),x=1} \frac{\exp\left(-\hat{\beta}_{12}\right)}{1-\hat{b}_{12}\exp\left(-\hat{\beta}_{12}\right)-\hat{b}_{13}\exp\left(-\hat{\beta}_{13}\right)}$$

$$\Rightarrow n_{12} = \sum_{\mathbf{I}(1,1),x=0}\left(\frac{\hat{b}_{12}}{1-\hat{b}_{12}-\hat{b}_{13}}\right) + \sum_{\mathbf{I}(1,1),x=1} \frac{b_{12}\exp\left(-\hat{\beta}_{12}\right)}{1-b_{12}\exp\left(-\hat{\beta}_{12}\right)-b_{13}\exp\left(-\hat{\beta}_{13}\right)}$$

$$= \sum_{\mathbf{I}(1,1),x=0}\left(\frac{\hat{b}_{12}}{1-\hat{b}_{12}-\hat{b}_{13}}\right) + X_{12} = n_{11,x=0}\left(\frac{\hat{b}_{12}}{1-\hat{b}_{12}-\hat{b}_{13}}\right) + X_{12}$$

$$\Rightarrow n_{12} = n_{11,x=0}\left(\frac{\hat{b}_{12}}{1-\hat{b}_{12}-\hat{b}_{13}}\right) + X_{12}$$

$$\Rightarrow \frac{n_{12} - X_{12}}{n_{11,x=0}} = \frac{\hat{b}_{12}}{1-\hat{b}_{12}-\hat{b}_{13}}$$

$$\hat{b}_{12} = \frac{n_{12} - X_{12}}{n_{11,x=0}}\left(1-\hat{b}_{12}-\hat{b}_{13}\right)$$

$$\Rightarrow \hat{b}_{12}\left(1+\frac{n_{12} - X_{12}}{n_{11,x=0}}\right) = \frac{n_{12} - X_{12}}{n_{11,x=0}}\left(1-\hat{b}_{13}\right)$$

$$\Rightarrow \hat{b}_{12} = \frac{\left(n_{12} - X_{12}\right)\left(1-\hat{b}_{13}\right)}{n_{11,x=0} + n_{12} - X_{12}}$$

$$n_{13} = n_{11,0}\left(\frac{\hat{b}_{13}}{1-\hat{b}_{12}-\hat{b}_{13}}\right)+X_{13}$$

$$\hat{b}_{13} = \frac{n_{13}-X_{13}}{n_{11,0}}\left(1-\hat{b}_{12}-\hat{b}_{13}\right) = \frac{n_{13}-X_{13}}{n_{11,0}}\left(1-\frac{\left(n_{12}-X_{12}\right)\left(1-\hat{b}_{13}\right)}{n_{11,x=0}+n_{12}-X_{12}}+\hat{b}_{13}\right)$$

$$\Rightarrow \hat{b}_{13}\left(1+\frac{n_{13}-X_{13}}{n_{11,0}}\right) = \frac{n_{13}-X_{13}}{n_{11,0}}\left(1-\frac{\left(n_{12}-X_{12}\right)\left(1-\hat{b}_{13}\right)}{n_{11,x=0}+n_{12}-X_{12}}\right)$$

$$\Rightarrow \hat{b}_{13}\left(\frac{n_{11,x=0}+n_{13}-X_{13}}{n_{11,x=0}}\right) = \frac{n_{13}-X_{13}}{n_{11,x=0}}\left(\frac{n_{11,x=0}-\hat{b}_{13}\left(n_{12}-X_{12}\right)}{n_{11,x=0}+n_{12}-X_{12}}\right)$$

$$\Rightarrow \hat{b}_{13}\left(\frac{n_{11,x=0}+n_{13}-X_{13}}{n_{11,x=0}}-\frac{\left(n_{12}-X_{12}\right)\left(n_{13}-X_{13}\right)}{n_{11,x=0}\left(n_{11,x=0}+n_{12}-X_{12}\right)}\right) = \frac{n_{13}-X_{13}}{n_{11,x=0}}\left(\frac{n_{11,x=0}}{n_{11,x=0}+n_{12}-X_{12}}\right)$$

$$\Rightarrow \hat{b}_{13}\left(\frac{n_{11,x=0}+n_{13}-X_{13}}{n_{11,x=0}}-\frac{\left(n_{12}-X_{12}\right)\left(n_{13}-X_{13}\right)}{n_{11,x=0}\left(n_{11,x=0}+n_{12}-X_{12}\right)}\right) = \frac{n_{13}-X_{13}}{n_{11,x=0}+n_{12}-X_{12}}$$

$$\Rightarrow \hat{b}_{13}\left(\left(n_{11,x=0}+n_{13}-X_{13}\right)\left(n_{11,x=0}+n_{12}-X_{12}\right)-\left(n_{12}-X_{12}\right)\left(n_{13}-X_{13}\right)\right) = \frac{\left(n_{13}-X_{13}\right)n_{11,x=0}}{n_{11,x=0}+n_{12}-X_{12}}$$

$$\hat{b}_{13} = \frac{\left(n_{13}-X_{13}\right)}{n_{12}-X_{12}+n_{13}-X_{13}+n_{11,x=0}} = \frac{n_{13,x=0}}{n_{12,x=0}+n_{13,x=0}+n_{11,x=0}}$$

$$\hat{b}_{13} = \frac{n_{13,x=0}}{n_{12,x=0}+n_{13,x=0}+n_{11,x=0}} \qquad \hat{b}_{12} = \frac{n_{12,x=0}}{n_{12,x=0}+n_{13,x=0}+n_{11,x=0}}$$

Finally we get all $\hat{p}_{ij}$ and $\hat{b}_{ij}$ then we can get all $\hat{\beta}$.

Intuitively the mle of transition probability is $\hat{P}_{ij} = \frac{x_{ij}}{x_{i.}}$.

We can divide data to subgroups and estimate all the transition probabilities in each subgroup. According to the invariance of mle, the function of mle will give a maximum likelihood estimator of the function. i.e $\hat{f}(\beta) = f(\hat{\beta})$

Under no such risk factor the transition probability is

$$\hat{P}_{ij,x=0} = \frac{x_{ij,x=0}}{x_{i.,x=0}}, \text{ and } \hat{P}_{ij,x=1} = \frac{x_{ij,x=1}}{x_{i.,x=1}}$$

For instance the model assumption is $P_{ij,x=0} = b_{ij}$, $P_{ij,x=1} = b_{ij} \exp(-\beta_{ij} x)$ On the other hand $b_{ij} = f(P_{ij,x=0}, P_{ij,x=0}) = f(\mathbf{P}_{ij})$ each transition probability is uncorrelated $\hat{b}_{ij} = f(\hat{\mathbf{P}}_{ij})$ and the same as $\hat{\beta}_{ij} = g(\hat{\mathbf{P}}_{ij})$

The multiple factors will use the same concept to derive.

2.3 The confidence intervals of parameter

Since finding a close form to the variance is different and not the only way to estimate the confidence interval of parameters. We try a simple method to evaluate the confidence interval of parameters. The method we used is so called **parametric bootstrap method** [4]. First estimate the mle of parameters and then we put all the factor history and initial state to the model. Generate the new sample and estimate the new sample mle then use all mle to develop the confidence intervals.

We trust the method will provide a reasonable variance. Bootstrap method has second order accuracy and however the Delta method is just first order accuracy.

In the beginning, the ideal is such that the following.

$$\mathrm{var}^*\left(\hat{\theta}\right)=\frac{1}{n^n-1}\sum_{i=1}^{n^n}\left(\hat{\theta}_i^*-\bar{\hat{\theta}}^*\right)^2$$

However if n is large then it is impossible to use the formula. Fortunately the new one is available.

$$\mathrm{var}_B^*\left(\hat{\theta}\right)=\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_i^*-\bar{\hat{\theta}}^*\right)^2 \quad B\text{: bootstrap sample size}$$

When n and B is large enough the bootstrap method has good properties.

$$\mathrm{var}_B^*\left(\hat{\theta}\right)\xrightarrow{B\to\infty}\mathrm{var}^*\left(\hat{\theta}\right) \quad \text{(Converge in probability)}$$

$$\mathrm{var}^*\left(\hat{\theta}\right)\xrightarrow{n\to\infty}\mathrm{var}\left(\hat{\theta}\right)$$

According to the new sample we can get a new mle. Sort all mle's to take 5_{th}% and 95_{th}% as the 90% CI.

More then one factors model the estimators are complex but we still can use the same concept to do.

To understand better the process and algorithm are like the following.

Process

1. Pick up predictors.
2. Construct the model.
3. Estimate the model parameters (maximum likelihood estimators).
4. Use the model, parameters, initial data and factor history to generate samples.
5. Go to 3 until the enough number of samples.
6. Order the parameters to construct the confidence intervals for each parameter.

Parametric bootstrap Algorithm

1. For some time t the transition probability with covariates $\boldsymbol{x}$ $[P^t_x]_{ij}$ (simply $p_{ij}(\boldsymbol{x})$). The state is from $S_1 \leqq i \leqq S_I$ and $S_1 \leqq j \leqq S_J$, where I=J
2. m=0, N subjects with their covariate classes x_m and initial state i_m for all $m<N$.
3. For $m<N$ generate a r.v. U_m or goto 5. The transition probability form stat i_m with covariate class $\boldsymbol{x}_m$ is $p_{ij}(\boldsymbol{x}_m)$. If U_m<summation of $p_{ij}(\boldsymbol{x}_m)$ wher j is from S_1 to j*+1 then assign $j_{m=}j$*
4. Estimate the mle_m then Go to 3
5. Sort all mle and get the 95% quintile and 5% quintile as the 90% CI.

3 Numerical method

3.1 Data exhibit

The Beaver Dam Eye Study is a longitudinal population based cohort study that aims at determining the long-term course of common vision-threatening conditions in adult Americans [9, 17]. Between September 15, 1987, and May 4, 1988, a private census was performed to identify residents of the city or township of Beaver Dam, Wisconsin, who were 43–84 years of age. A total of 5,924 persons were invited to participate in the study.

We pick up the all following participates. The total account is 927. Use these data to test homogeneity and stationarity. Then pick up the factors which are significant in above testing.

Table 3.1.1 The factor means or the count in different coding

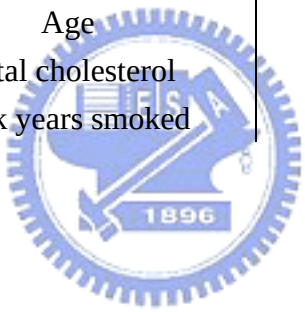
Factor	Baseline	5-year	10-year	15-year
Year of birth	1932.87	1932.87	1932.87	1932.87
Gender(1/0)	0.54	0.54	0.54	0.54
Age	55.43	60.22	65.48	70.25
hypertension(1/0)	0.28	0.36	0.46	0.58
Cholesterol	232.25	241	210.6	
History of drink	0.28	0.3	0.26	0.17
0/1/2	709/178/40	681/217/29	705/201/21	783/131/13
Smoke	1.32	1.25	1.19	1.17
0/1/2	20/587/320	17/665/245	18/719/190	0/770/157
Packages per year	27.96	28.92	29.24	29.16
Self reported vitamin use	0.88	1.17		
0/1/2	409/219/299	274/226/427		
3-level scale of worse eye (0/1/2)	779/144/4	737/183/7	716/200/11	677/226/24

Table 3.1.2 The codebook (for discrete data)

item\code	0	1	2
Gender	Female	male	x
Hypertension	normotensive	high BP	x
History of drink	Never	past	current
Smoke	Never	past	current
Vitamin supplement type	None	yes	yes (multivitamin)
Self reported vitamin use	Never	past	current

Table 3.1.3 The units in continuous data.

Item	Unit
Year of birth	year
Age	year
Total cholesterol	mg/dL
Pack years smoked	package



3.2 Results of numerical method

In order to simply the analysis we just select some categorical factors like smoke (the information of package of year included), *history of drinking* and *vitamin used* and two continuous factors *age* (if someone is older than 65) and *year of birth* (if someone is birth before than 1922) divided each into two subgroups. The data we use is all following in four times.

Table 3.2.1 Stationarity test

		stationarity					
	Subgroup	X=0	p-value	X=1	p-value	X=2	p-value
#2	Yob	3.477076	0.991136	5.25255	0.949007		
#3	Sex	6.182426	0.906608	4.491437	0.972851		
#4-7	Age	2.387382	0.998539	4.171494	0.980139		
#8-11	Hypten	9.453649	0.663772	3.101827	0.994781		
#15-18	hist	6.053536	0.913359	6.64387	0.880224	3.762936	0.987344
#19-22	smoke	1.780665	0.999675	4.837689	0.963182	12.53825	0.403476
#30-31	vithx	1.739159	0.999713	2.240092	0.998939	8.210214	0.768495

X is the risk factor code. X can be {0,1 and 2 }or just {0, 1}

Table 3.2.2 Homogeneity test

		Homogeneity					
	times	1	p-value	2	p-value	3	p-value
#2	yob	21.0670	0.00178	20.3829	0.00237	28.6282	0.00007
#3	sex	3.7381	0.71208	0.3694	0.99909	3.5203	0.74126
#4-7	age	18.2841	0.00556	18.2841	0.00556	23.7751	0.00057
#8-11	hypten	5.8456	0.44071	4.4790	0.61214	10.3536	0.11053
#15-18	hist	27.5529	0.00011	6.1372	0.40800	3.1723	0.78694
#19-22	smoke	2.7631	0.83794	13.4480	0.03645	9.3461	0.15503
#30-31	vithx	7.5825	0.27031	3.9208	0.68739	9.1054	0.16774

Table 3.2.3 the correlation of age and year of birth

	age	age2	age3
yob	-0.9953759	-0.9960287	-0.9956532

Although the *year of birth* is influential but *year of birth* and *age* are highly collinear. The different methods [7, 8] show the birth cohort effect but in this report we do not deal with both just select *age* in the model.

Finally, we select *the age (x_1)* and *history of drink (x_2)* in our model. The stationarity test is insignificant so we combine all the times under the Markov property assumption the next time a participant provides a independent data. In the assumption we believe that the risk factor would multiple the original probability then the statistical model is as following:

$$\mathbf{P} = [P_{ij}] = \begin{cases} P_{ij} = \exp\left(-\beta_{ij,0} - \beta_{ij,1}x_1 - \beta_{ij,2}x_2 - \gamma_{ij}x_1x_2\right), P_{02} = 0 \\ 1 - \Sigma_i \end{cases} \quad \forall i \neq j\,,\, i \neq 2$$

where $x_1, x_2 \in \{0,1\}$, and $\Sigma_i = \sum_{j \neq i} p_{ij}$

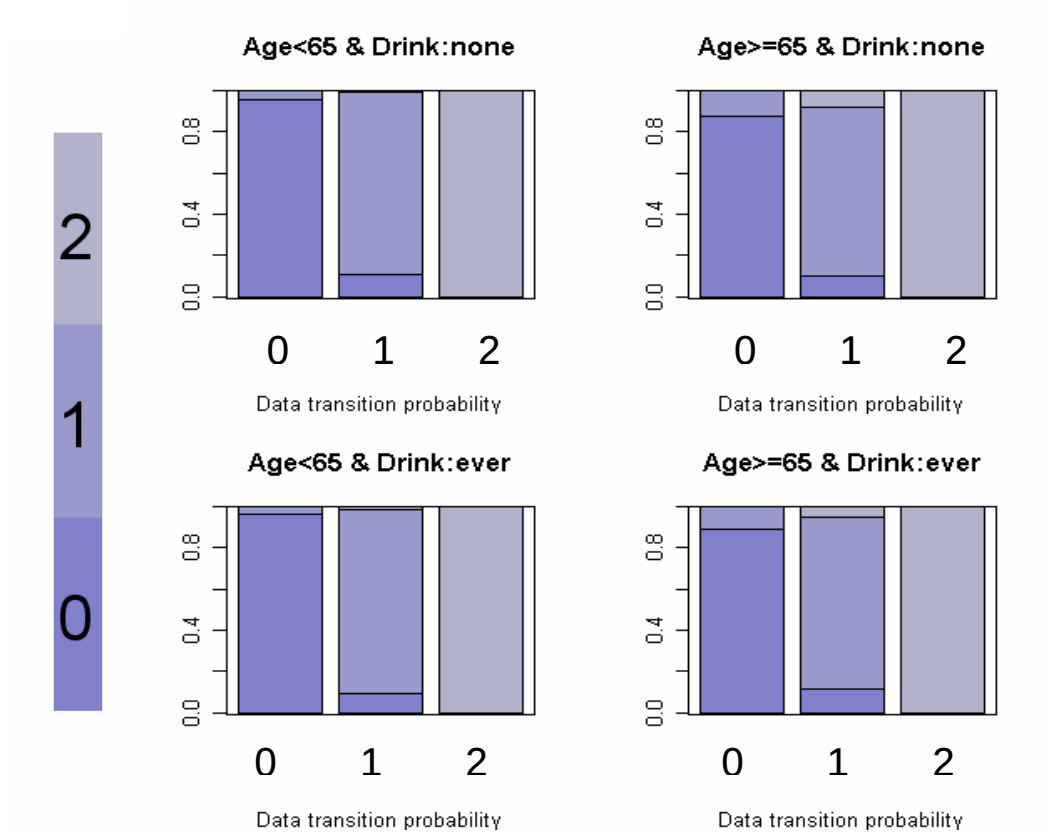


Figure 3.2.1 the transition probability of data

Figure 3.2.1 contains 4 graph illustrations of transition probabilities in each different covariate. The horizontal axes represent the starting state; the vertical axes represent the sum of the transition probability.

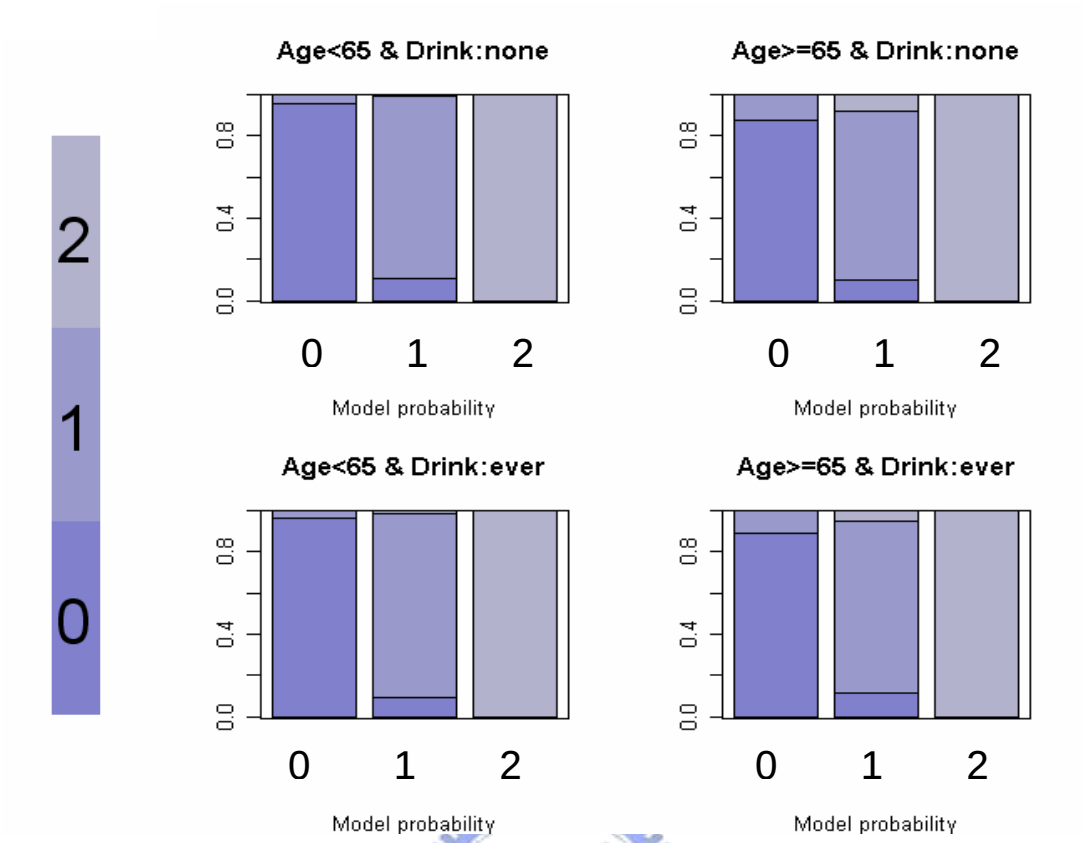


Figure 3.2.2 the transition probability of model

Figure 3.2.2 contains 4 graph illustrations of transition probabilities in each different covariate. The horizontal axes represent the starting state; the vertical axes represent the sum of the transition probability.

Table 3.2.4 mle of the parameters

	$\exp(-\hat{\beta}_0)$	$\exp(-\hat{\beta}_1)$	$\exp(-\hat{\beta}_2)$	$\exp(-\hat{\gamma})$
(0,1)	0.0462	2.7385	0.8354	1.0572
(1,0)	0.1105	0.9449	0.8653	1.2947
(1,2)	0.0103	8.2766	1.6913	0.3695

Table 3.2.5 parameter CIs using 100 rounds
each round has 5955 samples

	$\exp(-\hat{\beta}_0)$		$\exp(-\hat{\beta}_1)$		$\exp(-\hat{\beta}_2)$		$\exp(-\hat{\gamma})$	
	L	U	L	U	L	U	L	U
(0,1)	0.0394	0.0527	2.2240	3.2164	0.5743	1.1639	0.5662	1.8719
(1,0)	0.0874	0.1440	0.7217	1.3078	0.4228	1.3530	0.6861	2.6281
(1,2)	0.0026	0.0180	4.2324	34.6112	0.5638	10.1478	0.0439	1.2775

The underline represents significant.

Table 3.2.6 parameter CIs using 300 rounds
each round has 5955 samples

	$\exp(-\hat{\beta}_0)$		$\exp(-\hat{\beta}_1)$		$\exp(-\hat{\beta}_2)$		$\exp(-\hat{\gamma})$	
	L	U	L	U	L	U	L	U
(0,1)	0.0397	0.0527	2.2709	3.2750	0.5304	1.1166	0.6116	1.8667
(1,0)	0.0874	0.1362	0.6922	1.2958	0.4510	1.4160	0.6125	2.8056
(1,2)	0.0026	0.0206	4.0129	30.0967	0.5638	8.4565	0.0630	1.2305

The underline represents significant.

Table 3.2.7 parameter CIs using 1000 rounds
Each round has 5955 samples

	$\exp(-\hat{\beta}_0)$		$\exp(-\hat{\beta}_1)$		$\exp(-\hat{\beta}_2)$		$\exp(-\hat{\gamma})$	
	L	U	L	U	L	U	L	U
(0,1)	0.0397	0.0531	2.2604	3.3016	0.5338	1.1852	0.5939	1.7702
(1,0)	0.0823	0.1362	0.6824	1.3293	0.4510	1.4025	0.5698	2.8269
(1,2)	0.0026	0.0180	4.2637	32.3540	0.5638	6.7652	0.0595	1.1562

The underline represent significant.

3.3 Interpretation

Since the stationarity test is insignificant so we combine the different times to analysis and reduce the model to stationary model. Some probability is too small like (0,2) , (2,1), (2,2) so we assumption they are 0 in the model.

The testing data is used the all following that means no missing data in the data set and the total is 927 in three transient times . Once we select the factors then we used the no missing data just in the interesting factors so the number of data is increasing the total is 1985.

Before generate the CIs see **Figure 3.2.1** the transition probability plot show that the history of drink is insignificant by compare with the factor present or not. With the **Table** form **3.2.5** to **3.2.7** it is consist with the figure.

Some factors are effect in previous paper but they are not significant this. The reason may be some factors affect the ARM in part and the test is over all testing so they become not significant.

From the mle estimate we can see that the age will 2.7385 times from ARM free to early-ARM and 8.2766 times form early-ARM to late-ARM. Beside it seems to prevent from early-ARM to ARM free (0.9449 times but not significant).

In the common scene we may think the elder and the worse the ARM. In fact, *age* will not effect the disappear of ARM.

4 Conclusion

In this report we develop a new analysis of transition data with factors. Under the Markov property assumption we can easily solve the dependent data question but we need an “enough” sample size the better is uniform in each cell of probability. Use this method we will have a global view of a disease different from other methods in the past.

We use the ARM data to demonstrate the method in this example *age* happened to a factor so we do not need to develop a non-stationary model of course we can do it also.

The parametric model you can try any reasonable intuitively for factors and disease. It’s quite flexible.

The future work may try to develop a continuous time Markov chain with factors in the example five year maybe a little long.

5 Reference

1. A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E, beta carotene, and zinc for age-related macular degeneration and vision loss. Age-related Eye Disease Study Research Group. *Arch Ophthalmol* 2001;**119** 1417–36. (AREDS report no. 8).

2. Allain CC, Poon LS, Chan CGS, et al. Enzymatic determination of total serum cholesterol. *Clin Chem* 1974;**20** 470–5.

3. Basawa, Rao, Statistical Inference for Stochastic Process.

4. B. F. Efron, Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist*. 1979; **7** 1-26.

5. B. F. Efron, Computers and the Theory of Statistics: Thinking the unthinkable. *SIAM Review* . 1979; **21** 460-480.

6. George Casella, Roger L. Berger. Statistical Inference 2001

7. Huang GH, Ronald Klein, Barbara E. K. et al. Birth Cohort Effect on Prevalence of Age-related Maculopathy in the Beaver Dam Eye Study. *Am J Epidemiol* 2003;**157** 721–729

8. Huang GH, Klein R, Klein BEK: Joint analysis of incidence, progression, regression and disappearance rates. *Manuscript*

9. Klein R, Klein BEK, Linton KLP, et al. The Beaver Dam Eye Study: visual acuity. *Ophthalmology* 1991;**98** 1310–15.

10. Klein R, Klein BEK, Franke T. The relationship of cardiovascular disease and its risk factors to age-related maculopathy. The Beaver Dam Eye Study. *Ophthalmology* 1993;**100** 406–14.

11. Klein R, Klein BEK, Lee KP. The changes in visual acuity in a population. The Beaver Dam Eye Study. *Ophthalmology* 1996; **103** 1169–78.

12. Klein R, Klein BEK, Moss SE. Relation of smoking to the incidence of age-related maculopathy: the Beaver Dam Eye Study. *Am J Epidemiol* 1998; **147** 103–10.

13. Klein R, Klein BEK, Lee KE, et al. Changes in visual acuity in a population over a 10-year period: the Beaver Dam Eye Study. *Ophthalmology* 2001;**108** 1757–66.

14. Klein R, Klein BEK, Tomany SC, et al. Ten-year incidence of age-related maculopathy and smoking and drinking: the Beaver Dam Eye Study. *Am J Epidemiol* 2002; **156** 589–98.

15. Klein R, Klein REK, Wong TY, Tomany SC, Cruickshanks KJ. The association of cataract and cataract surgery with the long-term incidence of age-Related maculopathy. *Archives of Ophthalmology* 2002; **120** 1551-1558.

16. Klein R, Klein BEK, Tomany SC, et al. The association of cardiovascular disease with the long-term incidence of age-related maculopathy. The Beaver Dam Eye Study. *Ophthalmology* (in press).

17. Linton KLP, Klein BEK, Klein R. The validity of self-reported and surrogate-reported cataract and age-related macular degeneration in the Beaver Dam Eye Study. *Am J Epidemiol* 1991; **134** 1438–46.

18. S. Ross, Stochastic Process, 2nd Wiley.

19. The hypertension detection and follow-up program. Hypertension Detection and Follow-up Program Cooperative Group. *Prev. Med.* 1976; **5** 207–15.