The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.2.241
28(2), 241–250

Rank Tracking Probabilities using Linear Mixed Effect Models

Minjung Kwak[a,1]

[a]Department of Statistics, Yeungnam University

(Received March 16, 2015; Revised March 31, 2015; Accepted April 6, 2015)

Abstract

An important scientific objective of longitudinal studies involves tracking the probability of a subject having certain health condition over the course of the study. Proper definitions and estimates of disease risk tracking have important implications in the design and analysis of long-term biomedical studies and in developing guidelines for disease prevention and intervention. We study in this paper a class of rank-tracking probabilities to describe a subject's conditional probabilities of having certain health outcomes at two different time points. Linear mixed effects models are considered to estimate the tracking probabilities and their ratios of interest. We apply our methods to an epidemiological study of childhood cardiovascular risk factors.

Keywords: Longitudinal data, linear mixed effects model, rank tracking probability, rank tracking probability ratio.

1. 서론

경시적 자료(longitudinal data)는 각 개인에게서 관측치가 시간에 따라 반복적으로 얻어지는 경우에 발생한다. 연구에 참여한 각각의 개인들에 대해 시간의 흐름에 따라 규칙적으로 혹은 불규칙적으로 관측치가 얻어지며, 동일한 개체에서 관측치가 여러번 얻어지므로 관측치들이 서로 독립이라는 가정이 성립하지 않는다. 이러한 경시적 자료를 분석함에 있어서는 다변량 자료의 특성과 시계열 자료의 특성을 함께 고려하여야 한다. 첫째, 경시적 자료가 다변량 자료와 다른 특징은 관측치들이 시간에 따라 순서가 정해져 있다는 점이고, 둘째로 시계열 자료와 다른 특징은 시계열 자료와는 달리 한 개체에서 얻어지는 측정 시점들의 숫자가 상대적으로 적다는 점이다. 의학 통계에서 경시적 자료의 예로는 임상 시험에 있어서 두 가지의 서로 다른 치료법을 같은 환자에게 처리하여 반응변수의 변화를 관측하게 되는 교차설계에서 얻어지는 비교적 단순한 경시적 자료부터, 정기적으로 병원에 방문하여 동일한 환자에 대하여 각종 임상적 수치를 반복하여 관측, 기록하는 다소 복잡한 경시적 자료에 이르기까지 매우 다양하다.

경시적 자료에 대한 통계학적인 이론들은 Diggle 등 (1994), Liang 등 (2003)과 Lindsey (1993)에 의해 잘 정리가 되어 있고, 경시적 자료를 분석하기 위하여 사용된 각종 회귀 분석들에 대한 최근의 요약은

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2014R1A1A1002465).
[1]Department of Statistics, Yeungnam University, 280 Dae-Hak-Ro, Gyeongsan, Gyeongbuk 712-749, Korea. E-mail: mjkwak@yu.ac.kr

Molenberghs와 Verbeke (2005)에서 얻을 수 있다. 시간에 따라 반복적으로 관측치가 얻어지는 경시적 자료의 분석 목적은 여러 가지가 있으나, 그 중의 대표적인 분석 목적은 회귀 분석을 이용하여 관측된 반응변수와 독립변수간의 상관구조를 설명하거나 반응변수의 조건부 평균에 대한 추론을 하는 것이다.

본 연구의 배경이 된 자료는 미국의 한 연구기관에서 얻어진 아동들의 성장과 건강에 대한 경시적 자료이다. 이 자료는 사춘기 소녀들의 심혈관계 위험 요인에 대한 성향을 파악하고자 2500여명의 청소년기 소녀들을 대상으로 1986년부터 1997년까지 총 10번에 걸쳐 해마다 각종 임상 수치를 관측하여 기록한 자료이다. 수축기 혈압 또는 확장기 혈압이 각각 시간에 따라 반복 측정된 경우 이를 공변량의 함수로 설명하는 문제는 국내외적으로 중요한 연구 주제들 중의 하나이다. 본 자료의 임상 연구자들은 다른 독립 변수들이 어떤 수준으로 정해졌을 때의 혈압을 예측하는 문제와 어린 나이에 혈압이 높은 소녀들이 나중에 나이가 들었을 때도 계속 높은 혈압을 가지는 성향이 있는지를 판단하는 것에 관심이 있어 하였다. 이에 본 논문에서는 첫째 시간에 따라 반복 측정된 수축기 혈압이 관심 있는 공변량들의 함수로 설명하였으며, 둘째 특정 시점에서 혈압이라는 반응변수가 어떤 범위 안에 들어갈 때 그 이후의 다른 시점에서도 같은 범위 안에 들어가는지를 추적하는 측도인 순위 추적 확률과 순위 추적 확률비를 추정하고자 한다.

2. 대상 및 방법

2.1. 연구 대상

미국 국립보건원(NGHS)에서는 세 군데의 대학병원 센터들을 중심으로 1986년부터 2379명의 9–10세의 여아(백인 49%, 흑인 51%)들을 대상으로 1997년까지 10여년에 걸쳐 해마다 아동의 성장과 심혈관 질환의 위험요소를 측정하였다 (NHBPEP, 2004; NGHSRG, 1992). 지역은 전체 인구 중 인종별로 가구당 수입과 부모의 교육수준을 잘 반영할 수 있도록 선택되었다. 연구자들은 이 여아들을 대상으로 표준화된 프로토콜에 따라 부모의 동의를 거쳐 키, 몸무게, 혈압, 생활방식을 묻는 설문지 등을 해마다 측정하여 기록하였다. 추적율은 인종별로 백인 여아 74%에서 흑인 여아 95%에 이르며 여아들은 10년간 평균적으로 8.8번 센터를 방문하였다 (백인 8.6회, 흑인 9.0회). 혈압은 앉은 상태에서 표준화된 수은계를 이용하여 측정되었고, 키, 몸무게, 허리둘레 등은 표준화된 기구를 이용하여 실내에서 측정되었다. 실제로 자료가 얻어지는 과정에서는 i번째 개체가 j번째 시점에 병원을 방문하였을 때 임상 연구자가 여러번 혈압을 측정하여 그 측정치들의 평균을 자료로 기입하였다. 일반적인 심혈관계 질환의 위험요인 중 하나로 혈압을 고려하였으며, 수축기 혈압과 확장기 혈압 둘 중 어느 하나가 상위 95 분위수보다 큰 경우를 제1종 고혈압으로 정의하였고, 수축기 혈압과 확장기 혈압 둘 중 어느 하나가 상위 99 분위수보다 큰 경우를 제2종 고혈압으로 정의하였다 (Obarzanek 등, 2010). 키와 몸무게를 이용하여 BMI를 계산하고, BMI 가 상위 85 분위수보다 작은 경우 정상으로, 상위 85 분위수와 95 분위수 사이인 경우 과체중으로, 상위 95분위수보다 큰 경우 비만으로 정의하였다. 고혈압 유병율은 백인 여아의 경우 2.1%, 흑인 여아의 경우 5.0%였으며 비만인 여아의 경우 정상인 여아의 경우에 비해 고혈압 유병율은 6배 정도, 고혈압 발병율은 2–3배 정도 높은 것으로 분석되었다.

2.2. 자료 구조

우리는 경시적 자료 분석에서 일반적으로 쓰이는 다음의 자료 구조를 생각한다. 우리는 시간에 따라 반복 측정된 n명의 독립적인 개체들로 이루어진 경시적 자료를 가정하며 i번째 개체에서 n_i개의 관측치가 $t_{ij} \in \mathcal{T}$, $j = 1, \ldots, n_i$ 시점에서 얻어진다고 가정한다. 여기서 $\mathcal{T}$는 시간 인덱스를 나타내는 유계 집합으로 관심의 대상이 되는 관찰 기간을 나타낸다. 정해진 시점 $t \in \mathcal{T}$에 대하여, $Y(t)$는 실수 범위의 반

응변수를 나타내고, $\mathbf{X}(t) = (X_1(t), \ldots, X_p(t))^T$는 p 차원 실수 벡터로 표현되는 독립변수들의 집합이다. 윗첨자 T는 벡터나 행렬의 전치를 나타낸다. 경시적 자료 구조 $\{\mathbf{X}(t), Y(t), t)\}$에 대하여 연구자가 n명의 개체에 대하여 실제로 얻는 관측치는 $\{\mathbf{X}(t_{ij}), Y(t_{ij}), t_{ij};\ j = 1, \ldots, n_i,\ i = 1, \ldots, n\}$로 표현할 수 있다. 여기서 아래 첨자 i는 i번째 개체를 나타내며, 아래 첨자 j는 j번째 시점을 나타낸다. 여기서 유의할 점은 실제 임상 연구에서 개체를 연속적인 시간에 따라 관측하는 것은 불가능 하므로, n명의 각각의 개체는 서로 다른 관측 시점들의 집합 $\mathbf{t} = (t_1, \ldots, t_J)^T \in \mathcal{T}^J$, $J > 1$의 부분 집합에서 관측치가 얻어진다. 실제 연구에서 이상적인 상황은 미리 정해진 관측 시점에서 모든 개체들이 정확하게 관측되어지는 경우이지만, 임상 연구 진행에 있어 여러 가지 현실적인 문제로 미리 정해진 관측 시점에서만 관측치를 얻기는 거의 불가능한 상태이다. 다시 말하면 모든 개체들이 각각의 시점에서 관측되지는 않는다. 이 자료의 구조는 NGHS 자료의 구조와도 일치하는 것이다.

2.3. 통계 모형

선형 혼합 효과 모형(linear mixed effect model)은 의학, 공학, 심리학, 경제학 등 여러 학문 분야에서 광범위하게 사용되고 있는 통계 모형으로 시간에 따라 반복 측정된 경시적 자료의 분석에도 널리 쓰이고 있다. 본 논문에서는 여러 통계 모형들 중에서 선형 혼합 효과 모형을 NGHS 자료에 적용하여 소녀들의 성장에 따라 반복 측정된 수축기 혈압에 영향을 주는 요인들을 파악하고 수축기 혈압의 변화를 추적하는 확율을 추정하고자 한다. 선형 혼합 효과 모형은 각 개체에서 여러번 반복 측정된 종속 변수와 고정 효과(fixed effect) 변수와의 인과 관계 또는 상관관계를 설명하는데 많이 쓰이며 비선형 모형 또는 비모수 모형에 비해 해석이 용이하다는 이점을 가진다. n개의 독립적인 개체에서 얻은 자료를 바탕으로 한 선형 혼합 효과 모형의 일반적인 형태는 다음 식 (2.1)로 표현할 수 있다.

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \\ \mathbf{b}_i &\sim N(\mathbf{0}, \boldsymbol{\Psi}), \quad \boldsymbol{\epsilon}_i \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right), \end{aligned} \tag{2.1}$$

여기서 $\boldsymbol{\beta}$는 p 차원의 고정 효과 벡터이고, $\mathbf{b}_i$는 i번째 개체에 대한 q 차원의 확률 효과(random effect)의 벡터이며, $\mathbf{X}_i$와 $\mathbf{Z}_i$는 각각 i번째 개체에 대한 고정 효과와 확률 효과에 대응되는 디자인 행렬로, 각각 $n_i \times p$, $n_i \times q$ 차원이다. $\boldsymbol{\epsilon}_i$는 i번째 개체에 대한 개체 내 오차(within-subject error) 벡터를 나타낸다. 여기서 $\mathbf{b}_i$와 $\boldsymbol{\epsilon}_i$는 서로 독립이고 각각 평균이 0, 분산이 $\boldsymbol{\Psi}$, $\sigma^2\mathbf{I}$인 정규분포를 따른다. 확률 효과의 공분산 구조에 대하여 우리는 일반적인 양정치 행렬을 가정하였으며, 개체 내의 반복 요인이 확률 효과로 간주되었다.

한편 동일 개체의 오차라는 점을 고려하여 i번째 개체내의 오차 벡터 $\boldsymbol{\epsilon}_i$에 대하여 일반적인 공분산 행렬을 가지는 다변량 정규분포 $N(0, \sigma^2\boldsymbol{\Lambda}_i)$의 가정을 할 경우, 다음과 같은 모형식을 얻을 수 있다.

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \\ \mathbf{b}_i &\sim N(\mathbf{0}, \boldsymbol{\Psi}), \quad \boldsymbol{\epsilon}_i \sim N\left(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}_i\right). \end{aligned} \tag{2.2}$$

여기서 $\boldsymbol{\Lambda}_i$는 양정치 행렬이므로 $\boldsymbol{\Lambda}_i = (\boldsymbol{\Lambda}_i{}^{1/2})^T\boldsymbol{\Lambda}_i{}^{1/2}$로 쓰여질 수 있고, 모형식의 양 변에 $\boldsymbol{\Lambda}_i{}^{-1/2}$을 취함으로써 다음의 식을 얻을 수 있다.

$$\begin{aligned} \mathbf{y}_i^* &= \mathbf{X}_i^*\boldsymbol{\beta} + \mathbf{Z}_i^*\mathbf{b}_i + \boldsymbol{\epsilon}_i^*, \quad i = 1, \ldots, n, \\ \mathbf{b}_i &\sim N(\mathbf{0}, \boldsymbol{\Psi}), \quad \boldsymbol{\epsilon}_i^* \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right). \end{aligned} \tag{2.3}$$

여기서, $\mathbf{y}_i^* = (\boldsymbol{\Lambda}_i{}^{-1/2})^T\mathbf{y}_i$, $\mathbf{X}_i^* = (\boldsymbol{\Lambda}_i{}^{-1/2})^T\mathbf{X}_i$, $\mathbf{Z}_i^* = (\boldsymbol{\Lambda}_i{}^{-1/2})^T\mathbf{Z}_i$, 그리고 $\boldsymbol{\epsilon}_i^* = (\boldsymbol{\Lambda}_i{}^{-1/2})^T\boldsymbol{\epsilon}_i$이다. 그러므로 모형 (2.2)와 (2.3)은 본질적으로 같다고 볼 수 있다.

식 (2.1)의 모형에 따라 본 논문에서는 종속 변수로는 해마다 반복측정 된 수축기 혈압을 반응 변수로 사용하였고, 고정 효과 변수로는 인종, 나이, 키, 몸무게, BMI 백분위 수 등을 고려하였다. 확률 효과 변수로는 사용된 자료가 각 개체마다 10년간 해마다 반복 측정된 자료이므로 개체 요인을 확률 효과 변수로 하였다. 반응 변수 $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$을 이용한 선형 혼합 효과 모형의 우도 함수는 다음과 같다.

$$\begin{aligned} L\left(\boldsymbol{\beta}, \boldsymbol{\Psi}, \sigma^2 | \mathbf{y}\right) &= \prod_{i=1}^{n} p\left(\mathbf{y}_i | \boldsymbol{\beta}, \boldsymbol{\Psi}, \sigma^2\right) \\ &= \prod_{i=1}^{n} \int p\left(\mathbf{y}_i | \mathbf{b}_i, \boldsymbol{\beta}, \sigma^2\right) p\left(\mathbf{b}_i | \boldsymbol{\Psi}, \sigma^2\right) d\mathbf{b}_i, \end{aligned} \tag{2.4}$$

여기서 $\mathbf{y}_i$의 확률 밀도 함수는 다음과 같이 주어지며,

$$p\left(\mathbf{y}_i | \mathbf{b}_i, \boldsymbol{\beta}, \sigma^2\right) = \frac{\exp\left(-||\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i||^2/2\sigma^2\right)}{(2\pi\sigma^2)^{\frac{n_i}{2}}},$$

$\mathbf{b}_i$의 확률 밀도 함수는 다음과 같이 주어진다.

$$p\left(\mathbf{b}_i | \boldsymbol{\Psi}, \sigma^2\right) = \frac{\exp\left(-\mathbf{b}_i^T \boldsymbol{\Psi}^{-1} \mathbf{b}_i/2\right)}{(2\pi)^{\frac{q}{2}} \sqrt{|\boldsymbol{\Psi}|}}.$$

이렇게 설정된 선형 혼합 효과 모형의 모수값 $\boldsymbol{\beta}, \boldsymbol{\Psi}, \sigma^2$을 추정하기 위하여 위 (2.2)의 우도함수 또는 제한 우도함수를 최대화시키는 최대우도추정방법(Maximum Likelihood; ML)이나 제한 최대우도 추정방법(Restricted Maximum Likelihood; REML)을 사용한다.

2.4. 순위 추적 확률

본 연구에서는 관측된 경시적 자료를 대상으로 조건부 분포 함수와 순위 추적 확률에 대한 추론을 하고자 한다. 특정 시점 t에서 임의의 실수 부분 집합 A에 대하여 독립 변수가 $\mathbf{x}$로 주어졌을 때 반응 변수의 조건부 확률은 다음과 같이 주어진다.

$$P_A(\mathbf{x}, t) = \Pr\{Y(t) \in A(t) | \mathbf{X}(t) = \mathbf{x}(t)\}.$$

특별히 $A(t) = (-\infty, y(t)]$인 경우 위의 조건부 확률은 조건부 누적 분포 함수를 추정하는 문제가 된다. 조건부 누적 분포 함수에 대한 추론을 통해 조건부 분위수도 추론을 할 수가 있다. 임상 연구에 있어서 어떤 환자의 어렸을 적에 얻은 임상 관측치가 나이가 들어서 측정한 같은 관측치의 분포에 영향을 줄 수 있는지를 알아보기 위하여 추적 가능도(tracking ability)라는 개념을 도입하여 사용하고 있으며 두 관측시점 s_1과, s_2가 주어진 경우($s_1 < s_2$), 반응변수의 순위 추적 확률(rank-tracking probability; RTP)는 다음과 같이 정의된다.

$$\mathrm{RTP}_{s_1,s_2}(A, B) = \Pr\{Y(s_2) \in A(s_2) | Y(s_1) \in A(s_1), \mathbf{X}(s_1) \in B(s_1)\},$$

여기서 $A(s) \subset R$와 $B(s) \subset R^p$는 각각 미리 정해진 실수와 p 차원 실공간의 부분 집합들이다. RTP 값이 크면 독립 변수의 값이 s_1시점에서 B에 속하는 경우 반응 변수가 측정 시점 s_1과 s_2에서 집합 A에 속하는 추적 능력이 강하다고 해석한다. 또 다른 측도로 순위 추적 확률비(rank-tracking probability ratio; RTPR)가 있으며 이는 다음과 같이 정의된다.

$$\mathrm{RTPR}_{s_1,s_2}(A, B) = \frac{\mathrm{RTP}_{s_1,s_2}(A, B)}{\Pr\{Y(s_2) \in A(s_2) | \mathbf{X}(s_1) \in B(s_1)\}}.$$

Table 3.1. Esimates for the fixed effect parameters under the final model

고정 효과	추정값	표준 오차
Intercept	60.157	1.337**
Age	4.847	0.194**
Age^2	−0.140	0.007**
Race	0.739	0.249**
BMIPCT	0.094	0.003**

**는 1% 수준에서 통계적으로 유의함을 나타낸다.

$\text{RTPR}_{s_1,s_2}(A,B) = 1$인 경우, $Y(s_1) \in A(s_1)$임을 안다고 해서 $Y(s_2) \in A(s_2)$에 대한 확률이 증가하지는 않다는 것을 의미하며, 이는 $A(\cdot)$라는 집합에 대해 $Y(s)$가 추적 능력이 없다고 해석할 수 있다. 다른 한편으로 $\text{RTPR}_{s_1,s_2}(A,B) > 1$인 경우는 $Y(s)$가 양의 추적 능력을, $\text{RTPR}_{s_1,s_2}(A,B) < 1$인 경우는 $Y(s)$가 음의 추적능력을 가진다고 해석한다. 실제 임상 연구의 예로는 서로 다른 관측 시점에 대하여 두 시점 모두에서 고혈압군에 속하는 확률을 추적하는 능력을 평가하는 지표로 사용된다.

3. 결과

3.1. 선형 혼합 효과 모형

우리는 여러가지 독립 변수들의 조합과 확률 효과 모형을 고려한 여러 개의 선형 혼합 효과 모형을 적합하여 각 모형의 AIC, BIC, −2 Res Log likelihood 값을 비교하여 주어진 자료로 가장 적합이 잘 된 모형을 얻어내었다. 최종으로 선택된 수축기 혈압에 대한 선형 혼합 효과 모형의 분석 모형식은 다음과 같으며 최종 모형에 포함된 고정 효과에 대한 모수 추정값은 Table 3.1에 정리되었다.

$$y_{ij} = \beta_0 + \beta_1 \text{Age}_{ij} + \beta_2 \text{Age}_{ij}^2 + \beta_3 \text{Race}_i + \beta_4 \text{BMI}_{ij} + b_{i1} + b_{i2}\text{Age}_{ij} + \epsilon_{ij}. \tag{3.1}$$

각 개체별로 반복 측정에 의하여 얻어진 반응 변수를 설명하기 위하여 확률 효과 $\mathbf{b}_i = (b_{i1}, b_{i2})^T$를 고려하였으며 b_{i1}는 절편항에 대한 확률 효과를, b_{i2}는 Age에 대한 기울기항에 대한 확률 효과를 나타낸다. 여기에서 확률 효과 $b_i = (b_{i1}, b_{i2})^T$는 평균이 0이고 분산이 $\mathbf{\Psi} = \begin{pmatrix} \sigma_{b1}^2 & \rho\sigma_{b1}\sigma_{b2} \\ \rho\sigma_{b1}\sigma_{b2} & \sigma_{b2}^2 \end{pmatrix}$인 이변량 정규분포를 따른 다고 가정하였고, 오차항 ϵ_{ij}는 서로 독립이고 동일하게 정규분포 $N(0, \sigma^2)$를 따른다고 가정하였다.

최종으로 선택된 모형을 살펴보면, 나이에 대한 회귀 모수 추정치는 4.847 (p-value < 0.001)로, 나이가 증가함에 따라 수축기 혈압이 증가함을 볼 수 있다. 그러나 나이에 대한 이차항에 대한 회귀 모수 추정치가 −0.140 (p-value < 0.001)으로 나이의 증가에 따라 수축기 혈압이 계속적으로 증가하는 것이 아니라 10대 후반으로 감에 따라 수축기 혈압의 증가분이 감소하며 어느정 도 일정한 수준을 유지하는 패턴을 보이게 된다. 인종에 대한 고정 효과는 회귀 모수 추정치가 0.739 (p-value < 0.001)로서, 흑인 여아의 경우 백인 여아에 비해 수축기 혈압이 더 높음을 파악할 수 있다. 아동의 비만도를 판단하는 측도로 사용된 BMI 백분위수에 대한 회귀 모수 추정치는 0.094 (p-value < 0.001)로서 비만도가 높을수록 수축기 혈압이 작은 폭으로나마 증가함을 볼 수 있다. 각 개인마다 평균 10번의 관측치가 얻어졌는데, 이를 설명하기 위하여 포함한 i번째 개체에 대한 확률 효과 $b_i = (b_{i1}, b_{i2})^T$는 평균이 0이고 분산이 $\mathbf{\Psi}$인 정규분포를 따른 다고 가정하였다. 최종 모형에서 확률 효과의 분산 추정치는 $\widehat{\sigma}_{b1} = 9.115$, $\widehat{\sigma}_{b2} = 0.562$로 얻어졌으며, 개체 내 오차의 분산 추정치는 $\widehat{\sigma} = 6.168$로 얻어졌다. 이는 개체를 나타내는 확률 효과의 변동이 전체 변동의 약 69% ($= (\widehat{\sigma}_{b1}^2 + \widehat{\sigma}_{b2}^2)/(\widehat{\sigma}_{b1}^2 + \widehat{\sigma}_{b2}^2 + \widehat{\sigma}^2)$)를 차지하고 있으며, 개체에 의한 확률 효과의 변동이 상당 부분을 차지한다고 해석할 수 있다. 하지만 Age의 기울기항에 대한

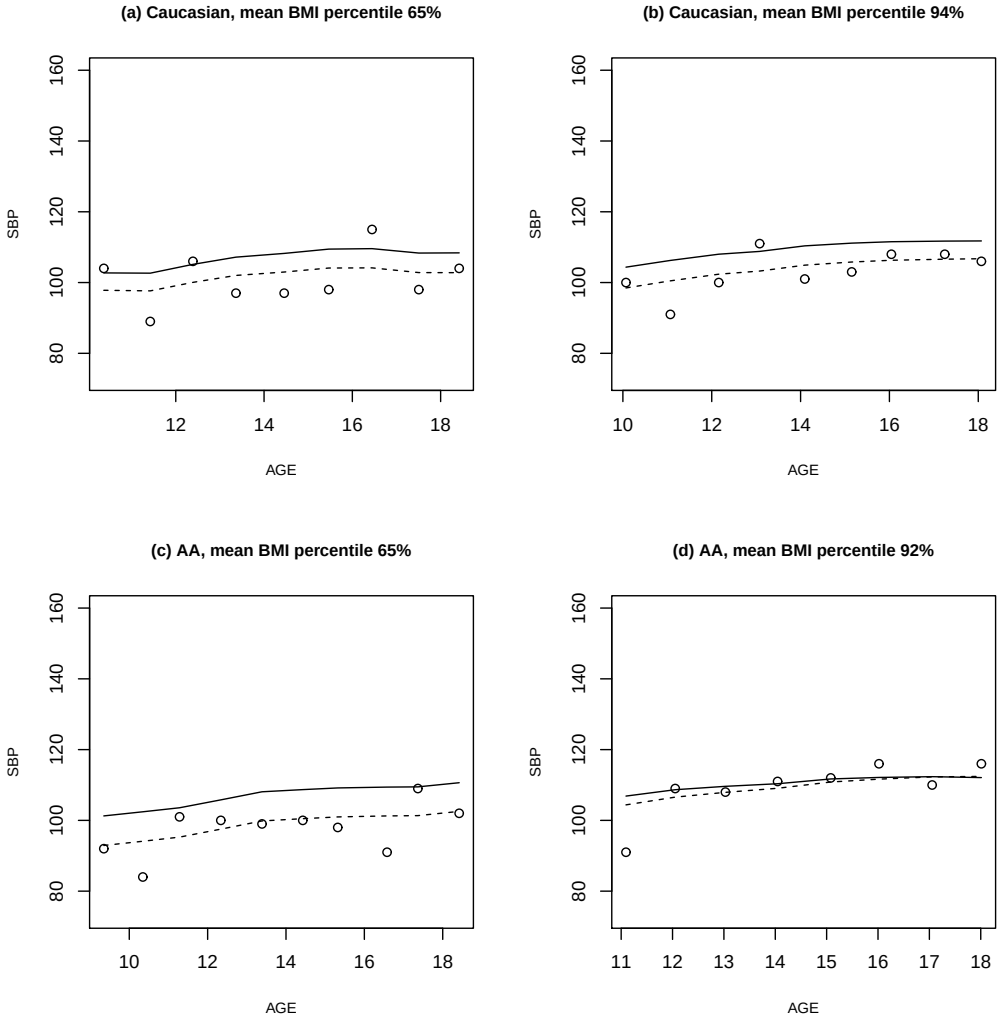

Figure 3.1. The longitudinal SBP measurements for four girls from NGHS with predicted subject-specific curves and mean population curves plotted in dashed and solid lines.

확률 효과의 변동이 전체 확률 효과의 변동에 미치는 영향은 미비한 것으로 보인다. 한편으로, 반복 측정된 혈압을 선형 혼합 효과 모형으로 설명하기 위해 가정한 개체 내 오차항에 대한 등분산 정규 분포의 가정이 성립하는 지를 살펴보기 위하여 개체 내 잔차의 정규확률지 그림을 검토하고 개체 내의 잔차를 적합값과 함께 그래프로 나타내어 보았을 때 0을 중심으로 고르게 퍼져있음을 확인하였다.

Figure 3.1는 2명의 백인 여아(BMI 백분위수 65%, 94%)와 2명의 흑인 여아(BMI 백분위수 65%, 92%)에 대하여 수축기 혈압의 원 자료와 평균 회귀 곡선, 그리고 개인별 수축기 혈압의 예측곡선을 나타낸다. 원 자료는 점으로 표시되었고, 평균 회귀 곡선은 실선으로 표시되었으며 개인 별 예측 곡선은 점선으로 표시되었다. 원 자료와 평균 회귀곡선, 개인별 예측곡선이 서로 가까이 배열되어 있음으로 보아 선택된 모형식 (3.1)이 합리적인 모형이라고 생각된다.

3.2. 순위 추적 확률

우리는 선택된 선형 혼합 효과 모형을 이용하여 순위 추적 확률과 순위 추적 확률비를 계산하였다. t시점에서 i번째 개체의 개인별 수축기 혈압의 예측값과 독립 변수를 각각 $\widetilde{Y}_i(t)$, $\mathbf{X}_i(t)$로 나타내었을 때, 서로 다른 두 시점 $s_1 < s_2$에서 반응 변수의 예측값이 A에 속하는 수축기 혈압의 순위 추적 확률(RTP)을 다음과 같이 구하였다.

$$\widehat{\mathrm{RTP}}_A(s_1, s_2) = \frac{\sum_{i=1}^{n} 1_{\left\{\widetilde{Y}_i(s_2)\in A(s_2), \widetilde{Y}_i(s_1)\in A(s_1), \mathbf{X}_i(s_1)\in B(s_1)\right\}}}{\sum_{i=1}^{n} 1_{\left\{\widetilde{Y}_i(s_1)\in A(s_1), \mathbf{X}_i(s_1)\in B(s_1)\right\}}}.$$

여기에서 1_A는 집합 A에 대한 표시 함수(indicator fuction)이다. 한편 위에 대응되는 순위 추적 확률비(RTPR)는 다음과 같이 구하였다.

$$\widehat{\mathrm{RTPR}}_A(s_1, s_2) = \frac{\widehat{\mathrm{RTP}}_A(s_1, s_2)}{\left(\frac{1}{n}\right)\sum_{i=1}^{n} 1_{\left\{\widetilde{Y}_i(s_1)\in A(s_1)\mathbf{X}_i(s_1)\in B(s_1)\right\}}}.$$

우리는 인종 별로 백인 여아, 흑인 여아에 대하여 수축기 확률이 120mmHg 이상(pre-hypertension)이 되는 순위 추적 확률과 순위 추적 확률비를 추정하였다. 즉, s_1시점에서 수축기 혈압이 120mmHg 이상인 경우, $s_2\,(> s_1)$시점에서 역시 수축기 혈압이 120mmHg일 순위 추적 확률을 계산하였다. Figure 3.2는 $s_2 = s_1 + 2$, 즉 특정 시점에 수축기 혈압이 120mmHg 이상인 경우 그 시점의 2년 후에도 수축기 혈압이 120mmHg 이상일 순위 추적 확률과 순위 추적 확률비에 대하여 그래프로 나타낸 것이다. Figure 3.2의 (a)와 (b)는 백인($x = 1$) 여아와 흑인($x = 2$) 여아에 대한 순위 추적 확률에 대한 그래프이다. 나이는 9세에서 14세까지를 고려하였다. 순위 추적 확률은 대략적으로 30%에서 40% 사이에 있으며, 백인 여아의 경우 흑인 여아에 비해 어렸을 적에 변동폭이 큼을 볼 수 있다. 평균적으로 백인 여아의 순위 추적 확률은 37.4% 정도였으며, 흑인 여아의 평균 순위 추적 확률은 32.3% 정도였다. Figure 3.2의 (c)와 (d)는 백인($x = 1$) 여아와 흑인($x = 2$) 여아에 대한 순위 추적 확률비에 대한 그래프이다. 평균적으로 백인 여아의 순위 추적 확률비는 5.44 정도였으며, 흑인 여아의 평균 순위 추적 확률비는 4.1 정도로 백인 여아의 경우 순위 추적 확률비가 흑인 여아에 비해 조금 큰 것을 확인할 수 있다. 순위 추적 확률과 마찬가지로 백인 여아의 경우 흑인 여아에 비해 어렸을 적에 추정치의 변동 폭이 큰 것은 볼 수 있다. 백인 여아, 흑인 여아 모두의 경우에서 1보다 큰 순위 추적 확률비를 보이므로 수축기 혈압은 양의 추적 능력을 가지는 임상 관측치라고 생각할 수 있다.

4. 결론

본 논문에서는 시간에 따라 반복 측정된 수축기 혈압 자료를 바탕으로 선형 혼합 효과 모형을 적용하여 자료를 설명하는 적절한 모형을 찾아내고, 순위 추적 확률과 순위 추적 확률비를 추정하였다. 순위 추적 확률은 어떤 개체의 건강 상태가 일정 시점에서 특정한 상태일 때 나중의 시점에서 같은 상태일 조건부 확률로 해석할 수 있다. 시점별 상관관계를 모형화하는 방법과는 달리 순위 추적 확률과 순위 추적 확률비는 실제 임상 연구에서 보다 간단하고 직접적인 해석이 가능하다는 장점이 있으며 시점별 상관관계를 모형화 하는데 필요한 가정들이 실제로 만족되는지에 대해 염려하지 않는다는 장점이 있다.

21세기 인류의 평균수명은 100세를 넘을 것이나 건강수명은 79세 밖에 되지 않아 약 20년을 심혈관계 질환 등과 같은 만성질환으로 고통 받을 것으로 추정되고 있다. 우리나라 통계청이 발표한 2007년 발표

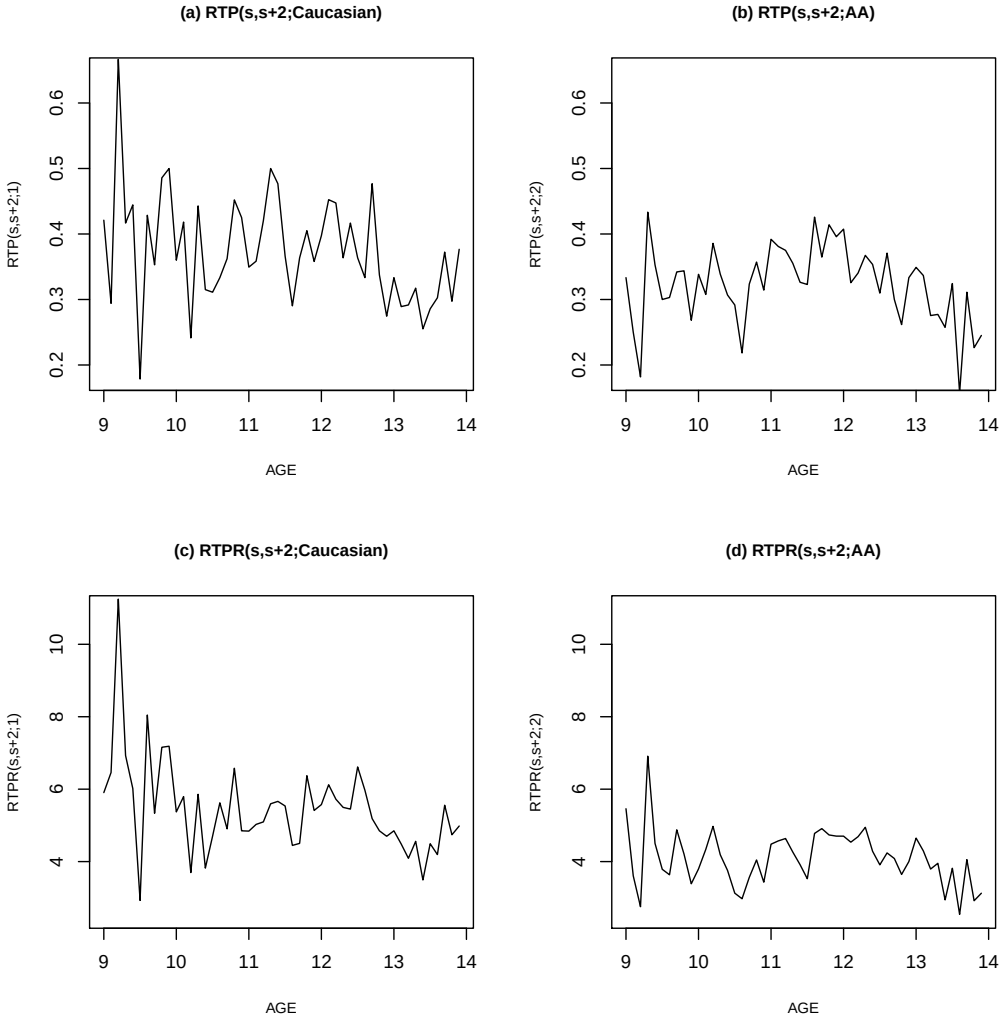

Figure 3.2. The estimated SBP RTP$(s, s+2; x)$ and RTPR$(s, s+2; x)$ for Caucasian ($x = 1$) and African-American ($x = 2$) girls.

한 보고서에 의하면 남성의 평균 수명이 75.1세, 여성이 82.3세로 2002년에 비해 각각 1.7세와 1.9세가 높아졌으며 평균적으로도 1.7세가 연장된 것으로 나타났다. 이에 따른 사망원인도 변화하여 생활습관성 질환의 사망률 비율은 점차 증가 추세로 40대까지는 심장질환으로 인한 사망률이 증가하며, 50대 이후로는 뇌혈관계질환으로 인한 사망률이 지속적인 증가세를 보이는 것으로 나타나 심혈관계 질환의 심각성이 부각되고 있다. 우리나라는 최근 인구의 고령화와 생활양식의 변화로 심혈관 질환이 급격히 증가하고 있는 추세이다. 특히 우리나라 30세 이상에서 고혈압과 고혈압 전기에 해당하는 비율이 58.5%에 달함에도 3–40대 성인 중 고혈압 환자는 대부분 본인이 환자라는 사실도 인지하지 못하고 있으며, 약물치료로 혈압을 적정 수준으로 유지하고 있는 환자 비율은 전체 환자 3명 중 1명꼴에 그치고 있다 (Korea Centers for Disease Control and Prevention, 2005). 본 연구에서는 10년간 경시적으로 얻어

진 혈압, 심혈관 위험요인 자료를 가지고 새로운 다변량 경시적 자료 분석 모형을 개발하고자 한다. 더욱이 제안된 모형을 통해 반복 측정으로 얻어진 다변량 자료에서 공변량의 효과를 추론해낼 뿐 아니라, 어떤 시점에서 고혈압군에 속했을 때 미래의 다른 시점에서 고혈압군에 속할 확률 등 현실적으로 자신의 혈압 상태를 추적하는데 도움이 되는 지표를 개발함으로써 임상연구에 있어서도 매우 유용하다.

본 연구는 심혈관계 질환의 위험요인을 파악하고자 수집된 임상연구 자료의 분석을 바탕으로 제안된 것이다. 실제적으로 심혈관계 질환 연구에 있어서 순위 추적 확률이나 순위 추정 확률비는 복합적인 병인을 지닌 심혈관계 질환 예방을 위해 중요하게 생각되고 실제적으로 많이 쓰이고 있는 척도이다. 본 연구에서 제안된 방법으로 위의 척도를 정확하게 추론해낸다면 실제 임상연구에도 활발히 적용이 될 수 있으며 노령인구가 증가하는 사회에서 고 연령대에서 많이 나타나는 심혈관계 질환을 이해하는 데에 중요한 역할을 할 수 있다.

References

Diggle, P. J., Liang, K. Y. and Zeger S. L. (1994). *Analysis of Longitudinal Data*, Oxford University Press, UK.

Korea Centers for Disease Control and Prevention (2007). Final analysis report of the third national health and nutrition examination survey, 35–64.

Liang, H., Wu, H. and Carroll, R. J. (2003). The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying coefficient models with measurement error, *Biostatistics*, **4**, 297–312.

Lindsey, J. K. (1993). *Models for Repeated Measurements*, Oxford University Press, UK.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*, Springer, New York.

National High Blood Pressure Education Program Working Group on High Blood Pressure in Children and Adolescents (NHBPEP Working Group) (2004). The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents. Pediatrics 114, 555– 576.

National Heart, Lung, and Blood Institute Growth and Health Research Group (NGHSRG) (1992). Obesity and cardiovascular disease risk factors in black and white girls: the NHLBI Growth and Health Study, *Am J. Public Health*, **82**, 1613–1620.

Obarzanek, E., Wu, C. O., Cutler, J. A., Kavey, R. W., Pearson, G. D. and Daniels, S. R. (2010). Prevalence and incidence of hypertension in adolescent girls, *The Journal of Pediatrics*, **157**, 461–467.

선형 혼합 효과 모형을 이용한 순위 추적 확률

곽민정[a,1]

[a]영남대학교 통계학과

(2015년 3월 16일 접수, 2015년 3월 31일 수정, 2015년 4월 6일 채택)

요 약

경시적 자료 연구의 중요한 주제 중의 하나는 시간이 지남에 따라 개인의 건강 상태가 어떻게 변하는지를 추적하는 확률이다. 질병의 상태를 시간의 흐름에 따라 추적하는 것은 장기간에 걸친 임상적 관찰 연구의 계획과 분석, 그리고 질병의 예방과 치료에 중요한 의미를 지닌다. 본 논문에서는 두 다른 시점에서 각 개인의 건강 상태에 대한 조건부 확률을 추정해내는 순위 추적 확률에 대하여 연구하였다. 순위 추적 확률과 순위 추적 확률비를 추정하기 위하여 선형 혼합 효과 모형을 고려하였다. 본 논문의 방법은 아동을 대상으로 심혈관계 질환의 위험요인을 연구하는 역학 자료에 적용되었다.

주요용어: 경시적 자료, 선형 혼합 효과 모형, 순위 추적 확률, 순위 추적 확률비.

이 논문은 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (NRF-2014R1A1A1002465).

[1](712-749) 경북 경산시 대학로 280, 영남대학교 통계학과. E-mail: mjkwak@yu.ac.kr