연세경제연구 제Ⅲ권, 제2호
Yonsei Economic Studies, Vol. Ⅲ, No. 2 (September 1996)

# 랜덤계수 회귀모형을 이용한 혈압강하제 효과분석

김병수 · 윤은영

본 연구에서는 반복측정된 시계열적 자료에 랜덤계수회귀모형을 사용하여 혈압강하제의 효과를 분석한다. 먼저 랜덤계수회귀모형에 대하여 개관하고, 고혈압 환자에게 혈압강하제를 투여한 후 90일 동안 추적하여 반복 관찰된 자료를 이 모형에 적합시킨다. 여기서 최대의 관심은 혈압강하제의 효과를 나타내는 전체평균화효과에 대한 추론이다.

이 모형은 관찰치가 서로 상관을 맺고 있으며, 특이한 공분산구조를 가진다. 독립된 개체는 시점(時点)에 따라 기울기가 각각 다르며, 이를 랜덤효과로 하여 주어진 모형에서 모수효과를 추정, 검정한다. 본 분석에서는 확장기혈압과 수축기혈압 자료 각각에서 공분산구조를 추정하였고, 이를 바탕으로 추정한 모수효과를 검정한 결과 고혈압 환자에게 투여한 혈압강하제는 혈압강하효과가 있는 것으로 판명되었다.

## Ⅰ. 서 론

본 연구에서는 혈압강하제를 환자에게 투여한 후 혈압강하효과를 알아보고자 실험한 반복측정(repeated measures)된 시계열적(longitudinal) 자료를 랜덤계수회귀모형(random coefficient regression model)을 이용하여 분석하고자 한다. 자료분석 목표는 시간이 흐름에 따라 혈압이 강하하는 여부와 정도를 살피는 것이다.

시계열적 자료는 시간 흐름에 따른 관찰치를 측정하며, 한 개체내에서 반복측정된 관찰치는 상관되어 있는 것이 특징이다. 반복측정이란 동일한 개체에서 실험조건, 처리를 달리하여 반응값을 얻거나 시점을 달리하여 반복적으로 값을 측정하는 경우이다. 랜덤계수회귀모형하에서 개인내 관찰치 사이에는 상관관계가 있고 개인내 변동부분이

연세대학교 응용통계학과, 서울특별시 서대문구 신촌동 134, 120-749. 한국 3M 물류본부, 서울특별시 영등포구 여의도동 27-3, 대한투자신탁빌딩 22층.

저자들은 항고혈압제의 임상연구 자료를 제공한 연세대학교 의과대학부속 세브란스병원 심장혈관 센터의 정남식 교수와 하종원 선생에게 감사를 표한다.

전체효과에서 일어나는 변동과 상관되어 있으며, 분산을 적절히 모형화시킬 때 랜덤계수회귀추정방법이 통상최소제곱(ordinary least square, 이하 OLS로 약칭함)추정보다 더 효율적이다 [11].

개인내 관찰치가 상관되어 있을 경우 OLS모형을 사용하여 모평균효과를 추정하면 추정치의 분산이 커지게 되어 정도(精度)가 떨어진다. 이와 같은 단점을 보완하기 위해서 시계열적 자료를 랜덤계수회귀모형에 적용시키고 이 모형에 대해 살펴보고자 한다.

이 연구에서는 반복측정된 시계열적 자료인 혈압 자료를 랜덤계수회귀모형에 적합시키는데 초점을 맞추었다. 제 Ⅱ 절에서는 이 모형에 대해 전반적으로 설명하였고, 그와 관련하여 모수효과(fixed effect)와 공분산구조를 덧붙였다. 제 Ⅲ 절에서는 모형의 모수를 추정하는 방법을, 제 Ⅳ 절에서는 EM알고리즘을 중심으로 해서 모수(parameter)를 추정, 계산하는 방법, 다른 접근법으로 일반화추정방정식(generalized estimating equation)과 깁스 샘플러(Gibbs sampler)를 언급하였다. 자료를 이용하여 모형을 설정하는 방법과 설정된 모형에서 잔차분석을 하는 문제는 제 Ⅴ 절에서 논의하였다. 마지막으로 제 Ⅵ 절에서는 혈압강하제를 고혈압 환자에게 투여한 후 반복측정된 혈압 자료를 이 모형에 적합시킨 후 혈압강하제의 효과를 분석하였다.

## Ⅱ. 랜덤계수회귀모형

### 1. 모 형

선형회귀모형에서 출발하여 랜덤계수회귀모형을 설명할 수 있다. 먼저 선형회귀모형을 제시하면 식 (1)과 같다.

$$Y_i = X_i\alpha + u_i, \qquad i = 1, 2, \cdots, m \tag{1}$$

$Y_i$ : 개체 $i$의 반응치를 나타내는 $n_i \times 1$벡터

$X_i$ : $n_i \times p$ 알려진 계획행렬

$\alpha$ : 추정해야 할 $p \times 1$ 모수벡터

$u_i \sim N(0, \sum_i)$이고 각각 독립

식 (1)의 선형회귀모형으로는 관찰된 종속변수들간에 상관관계를 설명할 수 없다. 종속변수들간의 상관관계를 모형화한 랜덤계수회귀모형을 구성하기 위하여 다음과 같이 가정한다.

$$u_i = Z_i b_i + e_i \quad (2)$$

$b_i \sim N(0, D)$, $e_i \sim N(0, R_i)$

$D$ : $k \times k$ 양정치(positive definite)공분산행렬

식 (1)에서 $u_i$의 분산은 $\sum_i = Z_i D Z_i' + R_i$로 나타내며, $Cov(b_i, e_i) = 0$, 즉 $b_i$와 $e_i$는 서로 독립이다. 식 (1)과 (2)를 사용하여 랜덤계수회귀모형을 다음의 식 (3)과 같이 나타낼 수 있다.

$$Y_i = X_i\alpha + Z_i b_i + e_i \quad (3)$$

이제 단계별로 나누어 다시 살펴보자. 모집단 모수, 개인효과와 개인내 변동은 단계 1에서, 개인간 변동은 단계 2에서 각각 도입된다.

### 단계 1

개체 각각에 대해 다음과 같다.

$Y_i = X_i\alpha + Z_i b_i + e_i$

$Z_i$ : $n_i \times k$ 알려진 계획행렬

$b_i$ : 추정해야 할 개체효과

$e_i$는 $N(0, R_i)$, 각각 독립이고 $R_i$는 $n_i \times n_i$ 양정치공분산행렬이다. 이 단계에서 $\alpha$와 $b_i$는 개체내에서 고정된 값으로 간주한다.

랜덤효과가 주어졌을 때 결과에 대한 조건부기대치함수는 알려지지 않은 변수와 공변수의 선형함수로 $h(E(y_i/b_i)) = X_i\alpha + Z_i b_i$이다. 그리고 $y$의 조건부공분산행렬은 $Var(y_i|b_i) = R_i$이다. 그러므로 $b_i$가 주어졌을 때 $y_i$의 조건부분포는 다음과 같다.

$$y_i|b_i \sim MVN(X_i\alpha + Z_i b_i, R_i) \quad (4)$$

### 단계 2

랜덤계수회귀모형의 2단계에서는 개체변동을 나타내는 랜덤효과분포 $\pi(b_i)$를 통해서 개인간 변동을 알 수 있다. $b_i$는 $N(0, D)$를 따르고 $i$ 각각에 대해 독립적이며 $e_i$와도 독립이다. 이 때 모수 $\alpha$는 모수효과로 취급한다. $f(y|b)$와 $\pi(b)$가 다변량정규분포를 한다면 $y$의 주변부분포는 다음의 식 (5)와 같다.

$$y_i \sim MVN(X_i\alpha, Z_i D Z_i' + R_i) \quad (5)$$

다변량정규조건부분포와 랜덤계수분포에 관련된 식 (2)와 같은 모형을 선형랜덤계

수회귀모형이라 한다. 일반화된 선형랜덤계수회귀모형(generalized linear random coefficient regression model)은 분포가 지수족(exponentail family)일 때를 말한다. 아주 특별한 경우를 제외하고 이 모형하에서는 $y_i$의 주변부분포를 구하기 어렵고 개체고유효과(subject specific effect)를 설명하기도 어렵다. 이러한 효과에 대해서는 제 Ⅱ.2소절에서 설명할 것이며, 이 연구에서는 선형랜덤회귀모형을 중점적으로 다룬다.

## 2. 모수효과

시계열적 자료에 대한 모형에서는 개체고유효과와 모집단평균화효과(population averaged effect)가 있다. 두 효과의 주된 차이는 나타내는 효과가 개인 반응치와 관련된 것인지 아니면 전체 반응과 관련된 것인지 하는 것이다. 즉, 개체고유효과에만 관심을 두는 모형과 전체평균화효과에만 역점을 두는 모형은 $X$가 변함에 따라 회귀계수가 개인의 반응을 나타내느냐 아니면 모집단의 평균화된 반응을 나타내느냐 하는데 차이가 있다. 개체고유효과는 기대되는 개인내 효과를, 모집단평균화효과는 $y$의 주변부기대값 $E_x(E(y_i|b_i))$에 근거하여 기대되는 개인간 효과를 모형화한다. 또 다른 차이는 반응치가 시간에 영향을 받는다는 사실에서 기인한다.

모집단평균화모형은 개체의 반복된 관찰치 사이의 공분산을, 개체고유모형은 이 공분산의 근원을 설명한다. $i$번째 개인의 개체고유계수는 $\alpha+b_i$이다. $E(b_i)=0$이므로 $\alpha$를 전형적인 개체고유모수로 볼 수 있으며, $E(b_i)=X_i\alpha$, $Cov(y_i)=Z_iDZ_i'+R_i$이므로 $\alpha$는 $X$가 공변량(共變量)이고 모집단반응이 $Y$일 때 변화비율로 해석할 수 있다. 또 선형랜덤계수회귀모형에서 $Y$의 주변부분포의 기대치를 이용하지 않고 주변부분포의 공분산행렬로 랜덤효과를 설명한다. 그러므로 선형랜덤계수회귀모형에서 백터 $\alpha$는 개체고유효과와 모집단평균화효과 모두를 나타낸다.

## 3. 분산구조

랜덤계수회귀모형에서는 특징적인 공분산구조를 볼 수 있다. 랜덤계수와 관련한 부분과 오차분산이 공분산 구조에 들어가며, 개체내 상관관계의 최소한 부분이 랜덤효과에 기인하는 것이다. 이 때 랜덤효과 분산행렬 $D$는 개인간 변동을, $Var(y_i|b_i)$인 $R_i$는 개인내 변동을 나타낸다. 그리고 분산행렬 $R_i$와 $D$는 모수백터 $\theta$의 알려진 함수로 가정한다.

$y_i$의 주변부분포의 분산, $Var(y_i)=Z_iDZ_i'+R_i$는 랜덤효과 공변수의 크기가 커지면 증가한다. 랜덤효과가 하나 이상일 때 관찰치 사이의 공분산도 랜덤효과 공변수가 커지면 증가하고, 공변수의 범위와 랜덤효과 사이의 상관관계에 영향을 받는다. 조건부공

분산 $R_i$는 공분산을 모형화한 것으로 랜덤효과 구조로 설명할 수 없다.

$y_i$의 주변부분포는 평균 $X_i\alpha$, 공분산행렬 $R_i+Z_iDZ_i'$ 인 다변량정규분포이다. $R_i=\sigma^2\cdot I$ 일 경우 이 모형을 조건부독립모형이라고 한다. 즉, $b_i$와 $\alpha$가 주어졌을 때 개체 $i$에 대한 $n_i$개의 반응치는 독립적이다. 선형랜덤계수회귀모형하에서 조건부독립은 등분산(homoscadastic variance)을 의미한다.

## Ⅲ. 추 정

### 1. 분산을 알 수 있는 경우

이 장에서는 모수 $\alpha$와 개체효과 $b_i$를 추정하며 그 분산을 구한다. 식 (3)의 모형에서 $y_i$의 공분산행렬 및 그 역행렬은 다음과 같이 표기한다.

$$V_i = Var(y_i) = R_i + Z_iDZ_i'$$
$$W_i = V_i^{-1}$$

그러면 $\alpha$와 $b_i$의 각각의 추정량 $\hat{\alpha}$과 $\hat{b}_i$은 다음의 식 (6), (7)과 같이 얻어진다.

$$\hat{\alpha} = (\sum_{i=1}^{m} X_i' W_i X_i)^{-1} \sum_{i=1}^{m} X_i' W_i y_i \tag{6}$$

$$\hat{b}_i = (\sum_{i=0}^{m} X_i' W_i X_i)^{-1} \sum_{i=1}^{n} X_i' W_i y_i \tag{7}$$

$\alpha$의 추정치는 자료의 주변부분포에 대한 우도를 최대화하였고 최소분산불편추정치(Uniformly Minimum Variance Unbiased Estimator)이다. $\hat{b}_i$의 표현은 최대우도는 아니며 가우스-마코프정리를 확장하여 얻어진 것이며 $\hat{b}_i = E(\hat{b}_i | y_i, \hat{\alpha}, \theta)$의 형태로 경험적 베이즈추정치(empirical bayes estimator)이다. $\hat{b}_i$을 $\hat{b}_i = DZ_i' W_i(y_i - X_i\hat{\alpha})$로 달리 표현할 수 있다. $\theta$는 $R_i$와 $D$ 내의 추정해야 할 분산-공분산 모수이다. 여기서 $b_i$의 사전평균(prior mean)이 0이므로 $\hat{b}_i$은 0과 $\overline{b}_i$의 가중결합이고 $\overline{b}_i$는 $b_i$를 모수효과로 취급해서 얻은 통상 가중최소제곱추정치(ordinary weighted least squares estimate)이다. 최대우도방법을 사용할 때 모수효과추정치의 근사적 분산은 역정보행렬 $I^{-1}(\alpha) = \{E(\partial^2 \log(L(\alpha, \theta; y)/\partial\alpha^2))\}^{-1}$을 이용하여 구한다. $\hat{\alpha}$과 $\hat{b}_i$는 $y$의 선형함수로 분산은 다음의 식 (8), (9), (10)과 같다.

$$Var(\hat{\alpha}) = (\sum_{i=1}^{m} X_i' W_i X_i)^{-1} \tag{8}$$

$$Var(\hat{b}_i) = DZ_i'(W_i - W_iX_i(\sum_{i=1}^{m} X_i' W_iX_i)^{-1}X_i' W_i)Z_iD \quad (9)$$

$$Var(\hat{b}_i - b_i) = D - DZ_i' W_iZ_iD + DZ_i' W_iX_i(\sum_{i=1}^{m} X_i' W_iX_i)^{-1}X_i' W_iZ_iD \quad (10)$$

### 2. 분산을 모르는 경우

$\hat{V}_i = \hat{R}_i + Z_i\hat{D}Z_i' = \hat{W}_i^{-1}$, $W_i$ 대신 $\hat{W}_i^{-1}$을 식 (6)과 식 (7)에 넣고 가중된 최소제곱방정식을 이용해서 $\alpha$와 $b_i$를 추정한다. 이 때 추정치 $\hat{\alpha}(\theta)$, $\hat{b}_i(\theta)$라고 표기하고 이것의 표준오차 추정치는 식 (8), (9)와 (10)에 $\hat{\theta}$를 넣고 구한다. $y$의 주변부분포에 대한 결합우도함수를 최대화시켜 $\alpha$와 $\theta$를 추정한다. 즉, 관찰된 자료의 우도 $L(\alpha, \theta: y) = \prod_{i=1}^{m} \int f(y_i : \alpha, \theta | b_i) \pi(b_i : \theta) db_i$를 사용하여 최대우도추정치(maximum likelihood estimation)를 구한다. 최대우도추정치를 직접 계산하는 연산으로는 제Ⅳ절에서 소개하는 EM연산법을 이용한다.

## Ⅳ. 계산 : EM연산법을 중심으로

### 1. EM연산법

여기서는 조건부독립랜덤회귀계수모형 $R_i = \sigma^2 \cdot I_{n_i \times n_i}$일 경우에 한해 언급하기로 한다. EM연산법의 M-단계와 E-단계를 각각 소개하면 다음과 같다.

[redacted]

$t$는 충분통계량이며 $t_1$, $t_2$는 추정된 충분통계량이고 $\theta$와 $t$의 최대우도추정치로 구의된 함수를 $M$이라 할 때, $\hat{\theta}$을 $t$의 함수로 나타내면 $\hat{\theta} = M(t)$이다. M-단계에서는 이를 이용하여 $\sigma^2$과 $D$에 대한 추정식을 구하면 식 (11), (12)가 된다.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{m} e_i' e_i}{\sum_{i=1}^{m} n_i} = \frac{t_1}{\sum_{i=1}^{m} n_i} \quad (11)$$

$$\hat{D} = m^{-1}\sum_{i=1}^{m} b_i b_i' = \frac{t_2}{m} \quad (12)$$

[redacted]

이 단계에서는 $t_1$과 $t_2$를 추정하고 그 식은 다음의 식 (13), (14)와 같다.

$$\hat{t}_1 = E\left[\sum_{i=1}^{m} e_i' e_i \mid y_i, \hat{\alpha}(\hat{\theta}), \hat{\theta}\right]$$

$$= \sum_{i=1}^{m} \left[e_i(\hat{\theta})' e_i(\hat{\theta}) + trVar(e_i \mid y_i, \hat{\alpha}(\hat{\theta}), \hat{\theta})\right] \quad (13)$$

$$\hat{t}_2 = E\left[\sum_{i=1}^{m} b_i b_i' \mid y_i, \hat{\alpha}(\hat{\theta}), \hat{\theta}\right]$$

$$= \sum_{i=1}^{m} \left[\hat{b}_i(\hat{\theta})\hat{b}_i(\hat{\theta})' + Var(b_i \mid y_i, \hat{\alpha}(\hat{\theta}), \hat{\theta})\right] \quad (14)$$

여기에서, $\hat{e}_i(\hat{\theta})$과 $\hat{t}$가 다음의 식 (15), (16)과 같이 얻어진다.

$$\hat{e}_i(\hat{\theta}) = E[e_i \mid y_i, \hat{\alpha}(\hat{\theta}), \hat{\theta}]$$

$$= y_i - X_i\hat{\alpha}(\hat{\theta}) - Z_i\hat{b}_i(\hat{\theta})$$

$$\hat{t} = E[t \mid y, \hat{\alpha}(\hat{\theta}), \hat{\theta}]$$

E-단계의 마지막 수렴단계에서 계산 결과 최대우도추정치를 구할 수 있다. 혼합모형(mixed model) 중 성장곡선모형(growth curve model)을 이용해 보자. 이는 EM알고리즘 단계에서 닫힌 형태(closed form)의 최대우도추정치를 설명하기 위해서이다.

완전한 자료이거나 균형 자료일 때, 즉 $n_i = n$이고 $Z_i = Z$일 때 모든 모수($\alpha$, $\theta$)에 대한 닫친 형태의 최대우도추정치가 존재하며 어떤 충분조건을 만족해야 한다. $y_i = X_i\alpha + Z_ib_i + e_i$를 아래와 같은 형태 식 (15)로 표현할 수 있어야 한다는 것이다.

$$Y = Z\Psi A + R \quad (15)$$

식 (15)는 제한적 성장곡선형태(restrictive growth curve form)이다.

$$y_i = Z_i\beta_i + e \qquad i = 1, 2, \cdots, m$$

$Z_i\beta_i$는 $i$번째 개인의 성장곡선이다. $\beta_i$는 각 개인의 랜덤모수(random parameter)로 개인마다 다르다.

$$\beta_i \sim N(A_i, D)$$

$$E(y_i) = Z_iA_i\alpha = X_i\alpha$$

$$Var(y_i) = \sigma^2 I_i + Z_iDZ_i'$$

$$X_i = Z_iA_i$$

$$\beta_i = A_i\alpha + b_i$$

$$r_i = y_i - X_i\alpha$$

$$r_i \sim N(0, \Sigma), \quad n_i = n$$

식 (15)에서 $R$은 $i$번째 열이 $r_i$인 $n \times m$행렬

$\Psi$는 $\alpha$를 적당히 재조정하여 얻은 $q \times r$행렬

$A$는 $i$번째 열이 $a_i$인 $r \times m$ 행렬

$\alpha^{(w)} = (\sum_{i=1}^{m} X_i' W_i^{(w)} X_i)^{-1} \sum_{i=1}^{m} X_i' W_i^{(w)} y_i$, $Z$는 $D$행렬의 형태와 상관없이 $\hat{\alpha} = \hat{\alpha}_{OLS} = (\sum_{i=1}^{m} X_i' X_i)^{-1} \sum_{i=1}^{m} X_i' y_i$로 축소할 수 있다. 임의의 공분산행렬을 갖는다는 가정 아래 $\Psi$의 최대우도추정치를 $\hat{\Psi}^*$로 표현하면 다음과 같다.

$$\hat{\Psi}^* = (Z' S^{-1} Z)^{-1} Z' S^{-1} Y A' (AA')^{-1}$$

$$S = Y(I - A'(AA')^{-1} A) Y$$

특히, $\sum = \sigma^2 I + ZDZ'$ 일 경우 $\hat{\Psi} = (Z'Z)^{-1} ZYA' (AA')^{-1}$이다.

공분산을 닫친 형태로 표현하면 다음과 같다.

$$\hat{\sigma}^2_{ML} = \frac{tr(Y' M_Z Y)}{(m(n-q))}$$

$$\hat{D}_{ML} = m^{-1} C_Z Y M_A Y' C_Z' - \hat{\sigma}^2_{ML} (Z'Z)^{-1}$$

$$M_Z = I_n - Z(Z'Z)^{-1} Z', \quad M_A = I_m - A'(AA')^{-1} A, \quad C_Z = (Z'Z)^{-1} Z'$$

## 2. 다른 접근 방법

여기에서는 EM연산법 외에 일반화추정방정식과 깁스 샘플러를 언급한다. 제 Ⅳ. 1소절에서 보인 것처럼 $\alpha$와 $D$의 추정식을 유도하지 않고 개념만을 간략하게 소개한다.

### 1) 일반화추정방정식

일반화추정방정식(generalized estimating equation)은 관찰치 주변부함수 $f(y_{ij})$가 지수족(exponential family)이라는 가정을 전제로 한다.

$$\sum_{i=1}^{m} \left( \frac{\partial E(y_{ij})}{\partial \alpha} \right) V_i^{-1}(\alpha) (y_i - E(y_{ij})) = 0 \qquad (16)$$

$V_i$ 는 모형화된 주변부공분산행렬 $Cov(y_i)$이다. 식 (16)을 통해서 모집단평균화효과 $\alpha$가 추정된다. $b_i = 0$일 때 대략 다음과 같다.

$$Cov(y_i) \approx Cov\left[ h^{-1}(X_{ti}\alpha) + \frac{\partial h^{-1}}{\partial b_i}(X_{ti}'\alpha)b_i \right] + \phi E\left\{ g[h^{-1}(X_{ti}\alpha) + \frac{\partial h^{-1}}{\partial b_i}(X_{ti}'\alpha)b_i] \right\}$$

$$\approx L_i Z_i D Z_i' L_i + \phi A_i = \hat{V}_i \qquad (17)$$

$$L_i = diag\left\{ \frac{\partial h^{-1}(u)}{\partial h}, u = X_{ti}'\alpha, t = 1, 2, \cdots, n_i \right\}$$

$Var(y_{ti}|b_i) = g(u_{ti}) \cdot \phi$로 나타낼 때 $g$는 연결함수(link function)이고 분산을 구하는 함수이며, $\phi$는 척도모수(scale parameter)이다. 식 (17)에서도 $\phi$는 척도모수이다. 이러한 방법을 써서 모수를 추정하며, 이에 대한 자세한 내용은 [9]와 [19]를 참조할 수 있다.

2) 깁스 샘플러

이 방법은 일반화선형랜덤계수회귀모형의 모수를 추정하는 방법 중 아주 융통성 있는 접근법이다. 베이즈 방법을 이용하므로 랜덤효과 분산 $D$에 대한 사전분포가 요구된다.

$$p(y_{ij}) = p(y_{ij}|\alpha, D, b_i) \cdot p(b_i|\alpha, D) \cdot p(\alpha, D) \qquad (18)$$

$p(\cdot)$는 확률분포함수이다.

주어진 $\alpha^{(0)}$, $b^{(0)}$, $D^{(0)}$에 대해, $\alpha^{(k+1)}$는 $p(\alpha \mid D^{(k)}, b^{(k)}, y)$에서, $D^{(k+1)}$는 $p(D \mid \alpha^{(k+1)}, b^{(k+1)}, y)$에서, $b^{(k+1)}$는 $p(b \mid \alpha^{(k+1)}, D^{(k+1)}, y)$에서 여러 번 반복하여 추출하면 $\alpha^{(*)}$, $b^{(*)}$, $D^{(*)}$의 분포는 결합사후확률분포함수(joint posterior p.d.f), $P(\alpha, b, D \mid y)$에 수렴한다. 그러나, 깁스 샘플링(Gibbs sampling)은 계층적 베이즈 모형(hierarchical Bayes model)을 정의하는 분포함수가 정확해야 한다. 제Ⅳ. 1소절에서와 마찬가지로 개념만을 간략하게 소개하였다. 이 방법은 근래에 많은 관심을 모으고 있으며 더욱 자세한 내용은 [14]를 참조할 수 있다.

## Ⅴ. 모형설정과 잔차분석

### 1. 모형설정

데이터를 분석해서 결과를 해석하기까지 몇 가지 단계로 나눌 수 있다.

첫 번째 단계로 모형을 선택한다. 분석하려는 자료와 연구문제에 타당하다고 판단되는 모형의 집합을 선택한다. 일반적으로 시계열적 자료를 랜덤계수회귀모형에 적합시킬 수 있다.

두 번째 단계에서는 모형설정에 있어서 가정을 점검해야 한다. 랜덤계수회귀모형을 살펴볼 때 개인내 상관관계는 개개인의 랜덤효과에서 기인한다는 것이 기본 가정이다. 이 가정의 타당성 여부는 랜덤효과를 포함하느냐에 달려 있다.

세 번째 단계는, 모형을 적합시키고 평가하는 작업이다. 주변부분산을 정확히 추정할 수 있으면 전체평균화효과의 효율성이 증가한다.

마지막 단계는 결과분석 단계이다. 선형랜덤계수회귀모형에서 모수효과는 개체고유효과와 전체평균화효과 모두를 모형화한다. 반면, 일반화선형랜덤계수회귀모형에서 모수효과는 개체고유효과만을 나타낸다.

### 2. 잔차분석

랜덤계수회귀모형에는 오차유형이 두 종류이다. 우선, 주변부오차를 $\epsilon_i$로 주변부잔차는 $e_i$로 표시하자. 주변부오차는 $\epsilon_i = y_i - \widehat{E(y_i)}$, 주변부오차추정치는 $\hat{\epsilon}_i = y_i - \widehat{E(y_i)}$이다. 주변부오차추정치, 즉 주변부잔차는 $e_i = y_i - E(y_i|b_i)$, 조건부오차추정치, 다시 말해서 조건부잔차는 $\hat{e}_i = y_i - E(y_i|b_i)$로 관찰치에서 추정된 모수효과와 랜덤효과를 제거하였다.

주변부잔차의 분산은 다음과 같다.

$$Var(\hat{\epsilon}_i) = V_i + X_i Var(\hat{\alpha}) X_i' - X_i Cov(\hat{\alpha}, y_i) - Cov(y_i, \hat{\alpha}) X_i'$$

만약, $Var(y_i)$를 알 수 있을 때 $\hat{\epsilon}_i$은 다변량정규분포를 하며 분산은 $Var(\hat{\epsilon}_i) = V_i - X_i Var(\hat{\alpha}) X_i'$ 이다. $V_i$는 주변부공분산행렬 $Cov(y_i)$이다. $\hat{\alpha}$의 추정치가 정확할수록 $Var(\hat{\alpha})$이 0에 가까운 경향이 있으며 $Var(\hat{\epsilon}_i)$은 $V_i$에 수렴한다.

조건부잔차는 주변부잔차의 선형함수 $\hat{e}_i = (I - Z_i D Z_i')\hat{\epsilon}_i$로 나타낼 수 있다. 그리고 그 분산은, $Var(\hat{e}_i) = V_i + X_i Var(\hat{\alpha}) X_i' + Z_i Var(\hat{b}_i) Z_i' - Cov(\hat{\epsilon}_i, \hat{b}_i) Z_i' - Z_i Cov(\hat{b}_i, \hat{\epsilon}_i)$이다. $V_i$, $D$와 $R_i$를 알 수 있을 때 $\hat{e}_i$는 다변량정규분포를 하며 분산은 $Var(\hat{e}_i) = (I - Z_i D Z_i')$, $Var(\hat{\epsilon}_i) = (I - Z_i D Z_i')$이다. $\hat{\alpha}$과 $\hat{b}_i$의 추정치가 정확할수록 모든 $i$에 대하여, $Var(\hat{\alpha})$가 0에, $Var(\hat{b}_i)$은 $D$에 근접해서 $Var(\hat{e}_i)$은 $R_i$에 수렴한다.

## Ⅵ. 반복측정된 혈압자료의 분석

실험대상은 연세대학교 의과대학부속 세브란스병원 심장혈관센터 심장내과 외래에 내원하여 본태성고혈압으로 진단 받은 사람으로서 확장기 혈압이 95mmHg에서 115mmHg이며 본 임상관찰에 동의한 환자이다.

확장기 혈압이 95mmHg 이상인 경우 랜덤하게 나누어 항고혈압제(cicletanine) 50mmHg이나 100mmHg을 투여하였는데 이전에 사용하였던 약제의 영향을 배제하기 위하여 충분한 시간 간격을 두고 실험을 행하였다. 이 과정에서 부작용을 보인 사람을 제외하고 각각 20명과 18명을 대상으로 하여 90일 동안 추적 관찰하였다.

항고혈압제 50mmHg과 100mmHg을 환자들에게 단독으로 투여하는 기간 동안 4주 간격-각 시점은 0, 4, 8, 12주-으로 앉은 자세와 기립자세에서 측정한 혈압을 반복치로 하였다.

통계패키지 BMDP [4]를 이용하여 제Ⅵ. 1소절에 정의되는 랜덤계수회귀모형에 적합시키면 결과는 제Ⅵ. 2소절과 제Ⅵ. 3소절과 같다.

## 1. 모 형

혈압자료에 랜덤계수회귀모형을 적용하기 위하여 식 (3)을 이용하자. 실제 적용할 모형은 다음과 같이 나타낼 수 있다.

$$y_{dit} = \alpha_{di} + b_{di}x_t + e_{dit} \tag{19}$$

식 (19)에서 지시자(indicator)가 나타내는 것은 다음과 같다.

$a$ : 50mmHg용량, 100mmHg용량을 나타내는 지시자

$i$ : 개인

$t$ : 시점(0주, 4주, 8주, 12주)

$\alpha_{di}$, $b_{di}$는 용량 $a$, 개인 $i$에 대한 랜덤절편과 기울기이고 $e_{dit}$는 랜덤오차 변동항이다. 그리고 각 항의 분포는 다음과 같다.

$$(\alpha_{di}, b_{di})' \sim N((\alpha_d, \beta_d)', D)$$
$$e_{di} = (e_{di1}, e_{di2}, e_{di3}, e_{di4})' \sim N(0, \sigma^2 \cdot I)$$
$$e_{dit} : a_{di}\text{와 } b_{di}\text{에 대해 독립}$$

이제, 이 모형을 아래와 같이 쓸 수 있다.

$$y_{di} = Z \cdot r_{di} + e_{di}$$
$$y_{di} = (y_{di1}, y_{di2}, y_{di3}, y_{di4})'$$
$$r_{di} = (\alpha_{di}, b_{di})'$$
$$Z = \begin{bmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 8 \\ 1 & 12 \end{bmatrix}$$

$y_{di}$의 공분산행렬은 랜덤효과구조, 즉 $\sum = ZDZ' + \sigma^2 \cdot I$이다.

## 2 확장기 혈압에 대한 결과 분석

### 1) 공분산 구조 추정

| | | |
|---|---|---|
| 1 | $d_{11}$ | 25.5380 |
| 2 | $d_{21}$ | .435016 |
| 3 | $d_{22}$ | .247720 |
| 4 | $\sigma^2$ | 21.5000 |

### 2) 개인내 공분산, 상관 행렬 ($\sum = ZDZ' + \sigma^2 \cdot I$)

$$\begin{bmatrix} 47.038047 & 27.278110 & 29.018173 & 30.758237 \\ 0.538846 & 67.342341 & 49.194195 & 48.907598 \\ 0.506238 & 0.627090 & 69.852395 & 58.019505 \\ 0.464670 & 0.623099 & 0.719269 & 93.150140 \end{bmatrix}$$

### 3) 회귀계수 추정치

| | | | | |
|---|---|---|---|---|
| *CONST* | 102.08566 | 0.72160 | 141.472 | 0.0000 |
| *dose*1 | −0.49066 | 0.72160 | −0.680 | 0.4965 |
| *x* | −0.33262 | 0.08140 | −4.086 | 0.0000 |
| *D*1.*X* | −0.12801 | 0.08140 | −1.573 | 0.1158 |

위 3)의 표에서 *dose*1항은 용량에 따른 절편의 추정치이며, *x*항은 용량 50mmHg, 100mmHg에서의 기울기 평균의 추정치이다. 그리고 *D*1.*x*항은 용량에 따른 기울기의 추정치이다.

위의 결과를 이용하여 회귀모수는 다음과 같이 계산된다

$$\hat{\alpha}_{50} = CONST + dose1 = 102.08566 + (-0.49066) = 101.595$$

$$\hat{\alpha}_{100} = CONST - dose1 = 102.08566 - (-0.49066) = 102.57632$$

$$\hat{\beta}_{50} = x + D1.x = -0.33262 - 0.12801 = -0.46063$$

$$\hat{\beta}_{100} = x - D1.x = -0.33262 - (-0.12801) = -0.20461$$

4) 모수효과와 공분산에 대한 유의성검정(Wald 검정)

| | | | |
|---|---|---|---|
| *dose* | 1 | 0.46 | 0.497 |
| *x* | 1 | 16.70 | 0.000 |
| *D*1.*x* | 1 | 2.47 | 0.116 |

왈드(Wald) 검정을 통해 알 수 있는 사실은 투약 50mmHg, 투약 100mmHg일 때 절편(dose항)과 기울기(*D.x*항)는 복용량에 따라 차이가 없다. 그러나, 위 검정에서 *x*는 투약 50mmHg과 100mmHg일 때 기울기의 평균은 0이라는 귀무가설을 검정하는 항이다. *P*-value < 0.05로 귀무가설을 기각하므로 시간에 따라 혈압강화효과가 있으며, 효과정도는 −0.33262 ± 0.08140이다.

## 3. 수축기 혈압에 대한 결과분석

1) 공분산 구조 추정

| | | |
|---|---|---|
| 1 | $d_{11}$ | 131.465 |
| 2 | $d_{21}$ | 1.31797 |
| 3 | $d_{22}$ | .618662 |
| 4 | $\sigma^2$ | 69.9449 |

2) 개인내 공분산, 상관 행렬 ($\sum = ZDZ' + \sigma^2 \cdot I$)

$$\begin{bmatrix} 201.40988 & 136.73690 & 142.00879 & 147.28068 \\ 0.64686 & 221.85226 & 167.07787 & 182.24836 \\ 0.61808 & 0.69288 & 262.09183 & 217.21604 \\ 0.57822 & 0.68174 & 0.74757 & 322.12859 \end{bmatrix}$$

3) 회귀계수 추정치

| | | | | |
|---|---|---|---|---|
| *CONST* | 152.63382 | 1.52141 | 100.324 | 0.0000 |
| *dose*1 | −4.36882 | 1.52141 | −3.872 | 0.0041 |
| *x* | −0.53587 | 0.13840 | −3.872 | 0.0001 |
| *D*1.*X* | 0.17400 | 0.13840 | 1.257 | 0.2087 |

위의 결과를 이용하여 회귀모수는 다음과 같이 계산된다

$$\hat{\alpha}_{50} = CONST + dosel = 152.63382 + (-4.36882) = 148.256$$
$$\hat{\alpha}_{100} = CONST - dosel = 152.63382 - (-4.36882) = 157.00264$$
$$\hat{\beta}_{50} = x + D1.x = -0.53587 + 0.17400 = -0.36187$$
$$\hat{\beta}_{100} = x - D1.x = -0.53587 - (-0.17400) = -0.70987$$

4) 모수효과와 공분산에 대한 유의성검정(Wald 검정)

| | | | |
|---|---|---|---|
| *dose* | 1 | 8.25 | 0.497 |
| *x* | 1 | 14.99 | 0.000 |
| *D.x* | 1 | 1.58 | 0.209 |

왈드(Wald) 검정을 통해 투약 50mmHg, 투약 10mmHg일 때 기울기(*D.x*항)는 복용량에 따라 차이가 없다는 사실을 알 수 있다. 그러나, 위 *dose*항과 *x*항을 보면 둘다 $P$-value < 0.05 로 귀무가설을 기각하므로 복용량에 따라 절편이 다르며, 시간에 따라 혈압강화효과가 있다. 그 효과 정도는 +0.53587 ± 0.13840이다.

## Ⅶ. 결 론

지금까지 랜덤계수회귀모형에 대해 전반적으로 알아보았고 실제 자료를 이 모형에 적용시켜 보았다. 앞서 밝힌 대로 이 모형은 개인내 관찰치들이 서로 상관되어 있으므로 특징적인 공분산구조를 보이게 된다. 그리고 이를 이용해서 모수(parameters)를 추정, 검정한다. 실제로 자료를 분석하는데는 EM연산법을 통해 모수를 추정하는 방법을 선택하고 확장기 혈압과 수축기 혈압에 대한 결과를 도출하였다.

이 연구에서는 랜덤계수회귀모형을 이용하여 실제로 반복측정된 혈압자료를 분석해 보았다. 근래에는 주변부공분산을 모형화하는데 랜덤효과와 자기상관오차모형(autoregressive error model)을 결합시킨다. 즉, 개체내 오차를 자기상관모형(autocorrelation model)과 병합시켜서 선형랜덤계수회귀로 모형화하는 부분에 관심이 높아져 가고 있는 실정이다.

## ◈참 고 문 헌◈


1. 하종원 · 임상욱 · 김병수 · 정남식 · 심원흠 · 조승연 · 김성순, "경증 및 중증도 고혈압 환자에서 Cicletanine 단독투여에 의한 강압효과 및 내약성 평가", 『순환기』, 제24권, 제3호, 1994, pp. 507~515.
2. Berkey, C. S. and Hoaglin, D. C. and Colditz, G. A., "A Random-Effects Regression Model for Meta-Analysis," *Statistics in Medicine*, 14, 1995, pp. 395~411
3. Blomquist, N., "On the Relation between Change and Initial Value," *Journal of American Statistical Association*, 72(360), 1977, pp. 746~749.
4. BMDP Statistical Software Manual, 1990, Berkeley, Los Angeles, Oxford, University of California Press.
5. Dempster, A. P., Laird, N. M. and Rubin, D. B., "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, B. 39, 1977, pp. 1~38.
6. Diggle, J. P., "An Approach to the Analysis of Repeated Measurements," *Biometrics*, 44, 1988, pp. 959~971.
7. Grizzle, J. E. and Allen, D. M., "Analysis of Growth and Dose Response Curves," *Biometrics*, 25, 1969, pp. 357~381.
8. Jennrich, R. I. and Schluchter, M. D., "Unbalanced Repeated Measures Models with Structured Covariance Matrices," *Biometrics*, 42, 1986, pp. 805~820.
9. Liang, K. Y. and Zeger, S. L., "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73(1), 1986, pp. 13~22.
10. Laird, N., Lange, N. and Stram, D., "Maximum Likelihood Computations with Repeated Measures: Application of the EM Algorithm," *Journal of American Statistical Association*, 82(397), 1987, pp. 97~105.
11. Laird, N. M. and Ware, J. H., "Random Effects Models for Longitudinal Data," *Biometrics*, 38, 1982, pp. 963~974.
12. Longford, T. N., "Random Coefficient Models," 1993, Clarendon Press, Oxford.
13. Louis, T. A., "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society*, B. 44(2), 1982, pp. 226~233.
14. Rao, C. R., "Simultaneous Estimation of Parameters in Different Linear Models and

Applications to Biometric Problems," *Biometrics*, 31, 1975, pp. 545~554.

15. Rutter, C. M. and Elashoff, R. M., "Analysis of Longitudinal Data : Random Coefficient Regression Modelling," *Statistics in Medicine*, 13, 1994, pp. 1211~1231.
16. Swamy, P. A. V. B., "Efficient Inference in a Random Coefficient Regression Model," *Econometrica*, 38(2), 1970, pp. 311~323.
17. Vonesh E. F. and Cater R. L., "Efficient Inference for Random Coefficient Growth Curve Models with Unbalanced Data," *Biometrics*, 43, 1987, pp. 617~628.
18. Wu, M., Ware, J. H. and Feinleib, M., "On the Relation between Blood Pressure Change and Initial Value," *Journal of Chronic Diseases*, 33, 1977, pp. 637~644.
19. Zeger, S. L., Liang, K. Y. and Albert, P. S., "Models for Longitudinal Data : A Generalized Estimating Equation Approach," *Biometrics*, 44, 1988, pp. 1049~1060.