Journal of the Korean Data & Information Science Society
2010, **21**(6), 1263–1270

한국데이터정보과학회지

콜레스테롤 자료에 대한 적정 공분산행렬 형태 산출에 관한 통계적 분석

조진남[1] · 백재욱[2]

[1]동덕여자대학교 정보통계학과 ·[2]한국방송통신대학교 정보통계학과

접수 2010년 10월 2일, 수정 2010년 11월 21일, 게재확정 2010년 11월 23일

요 약

60명의 환자들을 20명씩3개 그룹으로 나누어 각 그룹마다 다른 종류의 식이요법을 실시한 후 1주 간격으로 5주간에 걸쳐서 콜레스테롤 수치에 대한 반복측정 자료를 얻었다. 해당자료를 바탕으로 적합성여부와 유의성 검정을 실시한 결과 등분산 Toeplitz가 다양한 공분산행렬 형태들 중에서 가장 적합한 공분산구조 모형으로 판명되었다. 이 모형에서는 시점들 간의 상관계수는 0.64-0.78로 대체적으로 높은 상관관계들을 보여주고 있으며, 모수인자들의 유의성검정 결과, 시간효과는 대단히 유의하게 나타났으나, 처리 및 처리와 시간과의 교호작용효과는 유의하지 않은 것으로 판명되었다.

주요용어: 공분산행렬 형태, 모수인자, 반복측정자료, 우도함수, 혼합모형.

1. 서론

의학, 생물학 등에서 사람이나 동물들에 대하여 같은 개체를 대상으로 반복적으로 실험을 실시하여 얻어지는 자료를 반복측정자료라고 한다. 반복측정자료의 측정시점 간에는 시간별로 독립이 아닌 일정 관계의 연관성이 존재하며, 이러한 관계는 구체적인 공분산행렬 형태로 표현된다. 공분산행렬 형태에 대한 적절한 정보를 얻기 위해서는 혼합모형을 가정하여 분석하는 것이 모수모형 또는 변량모형으로 가정하여 분석하는 것보다 추정치의 분산이 적은 안정된 계수들을 얻을 수 있으며, 각 시점과의 연관성도 보다 잘 규명해 낼 수 있다. 혼합모형에 관련된 문헌들은 Brown과 Kempton (1994), Diggle (1989), Frees (2006), Hand와 Crowder (1996), Longford (1993), 조진남과 백재욱 (2009) 등이 있으며, 반복측정자료를 이용하여 공분산행렬 형태를 분석한 문헌으로는 Brown과 Prescott(1999), Fitzmaurice 등(2004), Verbeke와 Molenberghs (2000), 조진남 (2009) 등이 있다.

이 논문에서는 기존의 반복측정에 관련된 콜레스테롤 실험자료를 이용하여 다양한 공분산행렬 형태들 중에서 이 자료에 가장 적합한 공분산구조모형을 찾는 것이다. 이때 각 공분산행렬의 타당성은 모형의 적합성 여부와 유의성검정에 의하여 이루어지며, 가장 적합한 모형을 기준으로 모수인자들의 효과를 통계적으로 검정하고자 한다. 구체적으로 60명의 환자들을 대상으로 3종류의 식이요법을 각각 20명에게 투여한 후 5주간에 걸쳐 콜레스테롤에 관한 반복측정자료가 실험자료이며, 따라서 반응변수는 실험실시 후 1주 후마다 계속적으로 측정된 콜레스테롤 수치이며, 독립변수는 처리, 시점, 그리고 처리와 시점간의 교호작용이다

[1] (136-714) 서울시 성북구 월곡동 23-1, 동덕여자대학교 정보대학 정보통계학과, 교수.
[2] 교신저자: (110-791) 서울시 종로구 동숭동 169, 한국방송통신대학교 정보통계학과, 교수.
E-mail: jbaik@knou.ac.kr

2. 반복측정자료의 통계적 고찰

2.1. 혼합모형의 설정

n개의 관찰치가 주어졌을 때 반복측정자료의 혼합모형을 행렬형태로 표시하면 다음과 같다.

$$\underline{y} = X\underline{\alpha} + Z\underline{\beta} + \underline{e} \tag{2.1}$$

여기서 $\underline{y} = (y_1, y_2, \cdots, y_n)^{'}$는 $n \times 1$의 관찰치 벡터, X는 $n \times p$크기의 모수인자들의 계획행렬 (design matrix), $\underline{\alpha} = (\mu, \alpha_1, \alpha_2, \cdots, \alpha_p)^{'}$는 모수인자 계수 벡터, Z는 $n \times q$크기의 변량인자들의 계획행렬, $\underline{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^{'}$는 평균 $\underline{0}$, 분산공분산행렬 G인 다변량 정규분포를 하는 $q \times q$변량인자 계수 벡터, 그리고 $\underline{e}$는 평균 $\underline{0}$, 분산공분산행렬 R인 다변량 정규분포를 하는 $n \times 1$크기의 오차벡터이다. 이 때 $\underline{y}$의 분산공분산행렬 V는 다음과 같이 된다.

$$V = var(\underline{y}) = ZGZ^{'} + R \tag{2.2}$$

이 모형에서 모수인자 계수 $\underline{\alpha}$와 변량인자 계수 $\underline{\beta}$를 추정하기 위한 우도함수 (likelihood function)는 다음과 같다 (Brown과 Prescott, 1999).

$$L = (2\pi)^{-({}^1/{}_2)n} |V|^{-({}^1/{}_2)} \exp\left[-\frac{1}{2}(Y - X\alpha)^{'} V^{-1}(Y - X\alpha)\right] \tag{2.3}$$

이 우도함수를 최대화시키는 제한최우추정법 (REML: Restricted Maximum Likelihood)에 의하여 $\underline{\alpha}$와 $\underline{\beta}$는 다음과 같이 추정되며,

$$\widehat{\alpha} = \left(X^{'} V^{-1} X\right)^{-1} X^{'} V^{-1} \underline{y} \tag{2.4}$$

$$\widehat{\beta} = GZ^{'} V^{-1} (y - X\widehat{\alpha}) \tag{2.5}$$

추정량의 분산은 다음과 같다.

$$var(\widehat{\alpha}) = \left(X^{'} V^{-1} X\right)^{-1}$$
$$var\left(\widehat{\beta}\right) = GZ^{'} \left[V^{-1} - V^{-1}X(X^{'}V^{-1}X)^{-1}X^{'}V^{-1}\right] ZG$$

2.2. 반복측정자료의 공분산행렬 형태

$\underline{y}$의 분산공분산행렬 V는 식 (2.2)에서 보듯이 $ZGZ' + R$로 구성된다. $n \times n$ 크기의 오차의 공분산행렬 R은 각 개체내의 v개의 시점들 간의 공분산행렬 R_i을 대각항으로 표시하며, 비대각항은 개체들 간의 연관성이 존재하지 않으므로 0행렬로 이루어진다. 이 때 개체의 공분산행렬 R_i는 각 개체의 $v \times v$ 크기의 측정시점들 간의 공분산행렬이다. 따라서 전체 공분산행렬 R

$$R = \begin{bmatrix} R_1 & 0 & 0 & \cdots \\ 0 & R_2 & 0 & \cdots \\ 0 & 0 & R_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} = R_i \otimes I \tag{2.6}$$

로 표시할 수 있으며, 이는 개체의 공분산행렬 R_i가 어떤 형태를 가지느냐에 의해서 결정된다. 이 경우 개체내의 공분산행렬 R_i는 블록으로 간주되며, 개체내의 효과는 블록효과로 간주된다 그리고 개체내의 각 시점에 해당하는 관찰치 들은 특정 패턴을 가지는 것으로 가정한다.

R_i는 시점에 관계없이 분산값이 일정한 등분산 (homogeneous variance)인 경우와 시점 별로 분산이 다른 이분산 (heterogeneous variance)인 경우로 구분할 수 있다. 등분산과 이분산은 각각 단순형태 (simple), 복합대칭성 (confound symmetry), AR(1), Toeplitz(2), Toeplitz의 5가지 경우를 상정해 볼 수 있으며, 이외에도 멱형태(power), 지수형태 (exponential), 가우시안 (Gaussian), 일반적 형태 (unstructured) 등으로 구성된다.

이 논문에서는 등분산의 경우 단순형태, 복합대칭성, AR(1), Toeplitz(2), Toeplitz와 이분산의 경우 이분산 단순형태, 이분산 복합대칭성, 이분산 AR(1), 이분산Toeplitz(2), 이분산 Toeplitz의 총 10가지 경우를 상정하고, 이 실험자료에 가장 적합한 공분산행렬 형태를 찾고자 하며, 이에 따르는 처리, 시점, 그리고 처리와 시점간의 교호작용 효과를 검정하고자 한다. 전제조건으로 공분산행렬 형태는 처리 별로는 차이가 없는 것으로 가정한다. 각 모형의 구체적 형태는 표 3.1과 표 3.2에 정리하였다.

2.3. 모형의 적합성 검정과 유의성 검정

모형의 적합성 여부는 식 (2.3)의 우도값 기준으로는 그 값이 큰 모형이 작은 모형에 비해서 상대적으로 적합하다고 판단되며, $-2log(L)$와 AIC의 기준으로는 $-2log(L)$는 값이 작을수록, AIC는 값이 클수록 적합한 모형이라고 판단된다.

유의성검정은 어떤 공분산 형태의 모수가 다른 공분산형태의 일부분으로 포함되는 (Nested) 경우에 다른 공분산형태의 포함되지 않는 모수의 우도비검정 (LRT)을 이용하여 적합한 공분산구조 모형을 찾을 수 있다.

3. 콜레스테롤 실험자료 분석

3.1. 개체 공분산행렬 R_i의 추정

개체의 공분산 행렬 R_i는 등분산에 관련되는 5개의 공분산행렬과 이분산에 관련된 5개의 공분산행렬로 구분되며, 공분산행렬의 시점은 식이요법 실시 후 1주, 2주, 3주, 4주, 5주의 5개 시점이다. 이 실험자료를 바탕으로 등분산에 관련된 5개의 공분산행렬을 추정하여 표3.1에 수록하였으며, 이분산에 관련된 5개의 공분산행렬은 표3.2에 수록하였다. 자료분석은 혼합모형에 관련된 통계프로그램인 SAS의 PROC MIXED를 이용하였다 (Littell 등, 1996).

표 3.1 이분산의 공분산 행렬 형태 R_i와 추정결과

(i) 단순형태

$\hat{\sigma}^2 I = 2597I$

(ii) 복합대칭성

$$\hat{\sigma}^2 \begin{bmatrix} 1 & \hat{p} & \hat{p} & \hat{p} & \hat{p} \\ \hat{p} & 1 & \hat{p} & \hat{p} & \hat{p} \\ \hat{p} & \hat{p} & 1 & \hat{p} & \hat{p} \\ \hat{p} & \hat{p} & \hat{p} & 1 & \hat{p} \\ \hat{p} & \hat{p} & \hat{p} & \hat{p} & 1 \end{bmatrix} = 2597 \begin{bmatrix} 1 & 0.7324 & 0.7324 & 0.7324 & 0.7324 \\ 0.7324 & 1 & 0.7324 & 0.7324 & 0.7324 \\ 0.7324 & 0.7324 & 1 & 0.7324 & 0.7324 \\ 0.7324 & 0.7324 & 0.7324 & 1 & 0.7324 \\ 0.7324 & 0.7324 & 0.7324 & 0.7324 & 1 \end{bmatrix}$$

표 3.1 (계속) 이분산의 공분산 행렬 형태 R_i와 추정결과

(iii) AR(1)

$$\hat{\sigma}^2\begin{bmatrix}1 & \hat{p} & \hat{p}^2 & \hat{p}^3 & \hat{p}^4\\ \hat{p} & 1 & \hat{p} & \hat{p}^2 & \hat{p}^3\\ \hat{p}^2 & \hat{p} & 1 & \hat{p} & \hat{p}^2\\ \hat{p}^3 & \hat{p}^2 & \hat{p} & 1 & \hat{p}\\ \hat{p}^4 & \hat{p}^3 & \hat{p}^2 & \hat{p} & 1\end{bmatrix} = 2713\begin{bmatrix}1 & 0.7703 & 0.5933 & 0.4570 & 0.3520\\ 0.7703 & 1 & 0.7703 & 0.5933 & 0.4570\\ 0.5933 & 0.7703 & 1 & 0.7703 & 0.5933\\ 0.4570 & 0.5933 & 0.7703 & 1 & 0.7703\\ 0.3520 & 0.4570 & 0.5933 & 0.7703 & 1\end{bmatrix}$$

(iv) Toeplitz(2)

$$\hat{\sigma}^2\begin{bmatrix}1 & \hat{p}_1 & 0 & 0 & 0\\ \hat{p}_1 & 1 & \hat{p}_1 & 0 & 0\\ 0 & \hat{p}_1 & 1 & \hat{p}_1 & 0\\ 0 & 0 & \hat{p}_1 & 1 & \hat{p}_1\\ 0 & 0 & 0 & \hat{p}_1 & 1\end{bmatrix} = 2265\begin{bmatrix}1 & 0.4171 & 0 & 0 & 0\\ 0.4171 & 1 & 0.4171 & 0 & 0\\ 0 & 0.4171 & 1 & 0.4171 & 0\\ 0 & 0 & 0.4171 & 1 & 0.4171\\ 0 & 0 & 0 & 0.4171 & 1\end{bmatrix}$$

(v) Toeplitz

$$\hat{\sigma}^2\begin{bmatrix}1 & \hat{p}_1 & \hat{p}_2 & \hat{p}_3 & \hat{p}_4\\ \hat{p}_1 & 1 & \hat{p}_1 & \hat{p}_2 & \hat{p}_3\\ \hat{p}_2 & \hat{p}_1 & 1 & \hat{p}_1 & \hat{p}_2\\ \hat{p}_3 & \hat{p}_2 & \hat{p}_1 & 1 & \hat{p}_1\\ \hat{p}_4 & \hat{p}_3 & \hat{p}_2 & \hat{p}_1 & 1\end{bmatrix} = 2641\begin{bmatrix}1 & 0.7609 & 0.7750 & 0.6931 & 0.6400\\ 0.7609 & 1 & 0.7609 & 0.7750 & 0.6931\\ 0.7750 & 0.7609 & 1 & 0.7609 & 0.7750\\ 0.6931 & 0.7750 & 0.7609 & 1 & 0.7609\\ 0.6400 & 0.6931 & 0.7750 & 0.7609 & 1\end{bmatrix}$$

표 3.2 이분산의 공분산 행렬 형태 R_i와 추정결과

(vi) 이분산 단순형태

$$\begin{bmatrix}\hat{\sigma_1}^2 & 0 & 0 & 0 & 0\\ 0 & \hat{\sigma_2}^2 & 0 & 0 & 0\\ 0 & 0 & \hat{\sigma_3}^2 & 0 & 0\\ 0 & 0 & 0 & \hat{\sigma_4}^2 & 0\\ 0 & 0 & 0 & 0 & \hat{\sigma_5}^2\end{bmatrix} = \begin{bmatrix}2723 & 0 & 0 & 0 & 0\\ 0 & 2082 & 0 & 0 & 0\\ 0 & 0 & 2691 & 0 & 0\\ 0 & 0 & 0 & 2390 & 0\\ 0 & 0 & 0 & 0 & 3098\end{bmatrix}$$

(vii) 이분산 복합대칭성

$$\begin{bmatrix}\hat{\sigma_1}^2 & \hat{p}\hat{\sigma_1}\hat{\sigma_2} & \hat{p}\hat{\sigma_1}\hat{\sigma_3} & \hat{p}\hat{\sigma_1}\hat{\sigma_3} & \hat{p}\hat{\sigma_1}\hat{\sigma_5}\\ \hat{p}\hat{\sigma_1}\hat{\sigma_2} & \hat{\sigma_2}^2 & \hat{p}\hat{\sigma_2}\hat{\sigma_3} & \hat{p}\hat{\sigma_2}\hat{\sigma_4} & \hat{p}\hat{\sigma_2}\hat{\sigma_5}\\ \hat{p}\hat{\sigma_1}\hat{\sigma_3} & \hat{p}\hat{\sigma_2}\hat{\sigma_3} & \hat{\sigma_3}^2 & \hat{p}\hat{\sigma_3}\hat{\sigma_4} & \hat{p}\hat{\sigma_3}\hat{\sigma_5}\\ \hat{p}\hat{\sigma_1}\hat{\sigma_4} & \hat{p}\hat{\sigma_2}\hat{\sigma_4} & \hat{p}\hat{\sigma_3}\hat{\sigma_4} & \hat{\sigma_4}^2 & \hat{p}\hat{\sigma_4}\hat{\sigma_5}\\ \hat{p}\hat{\sigma_1}\hat{\sigma_5} & \hat{p}\hat{\sigma_2}\hat{\sigma_5} & \hat{p}\hat{\sigma_3}\hat{\sigma_5} & \hat{p}\hat{\sigma_4}\hat{\sigma_5} & \hat{\sigma_5}^2\end{bmatrix} = \begin{bmatrix}2665 & 1750 & 1940 & 1846 & 2209\\ 1750 & 2102 & 1722 & 1639 & 1961\\ 1940 & 1722 & 2583 & 1817 & 2174\\ 1846 & 1639 & 1817 & 2340 & 2070\\ 2209 & 1962 & 2175 & 2070 & 2070\end{bmatrix}$$

(viii) 이분산 AR(1)

$$\begin{bmatrix}\hat{\sigma_1}^2 & \hat{p}\hat{\sigma_1}\hat{\sigma_2} & \hat{p}^2\hat{\sigma_1}\hat{\sigma_3} & \hat{p}^3\hat{\sigma_1}\hat{\sigma_3} & \hat{p}^4\hat{\sigma_1}\hat{\sigma_5}\\ \hat{p}\hat{\sigma_1}\hat{\sigma_2} & \hat{\sigma_2}^2 & \hat{p}\hat{\sigma_2}\hat{\sigma_3} & \hat{p}^2\hat{\sigma_2}\hat{\sigma_4} & \hat{p}^3\hat{\sigma_2}\hat{\sigma_5}\\ \hat{p}^2\hat{\sigma_1}\hat{\sigma_3} & \hat{p}\hat{\sigma_2}\hat{\sigma_3} & \hat{\sigma_3}^2 & \hat{p}\hat{\sigma_3}\hat{\sigma_4} & \hat{p}^2\hat{\sigma_3}\hat{\sigma_5}\\ \hat{p}^3\hat{\sigma_1}\hat{\sigma_4} & \hat{p}^2\hat{\sigma_2}\hat{\sigma_4} & \hat{p}\hat{\sigma_3}\hat{\sigma_4} & \hat{\sigma_4}^2 & \hat{p}\hat{\sigma_4}\hat{\sigma_5}\\ \hat{p}^4\hat{\sigma_1}\hat{\sigma_5} & \hat{p}^3\hat{\sigma_2}\hat{\sigma_5} & \hat{p}^2\hat{\sigma_3}\hat{\sigma_5} & \hat{p}\hat{\sigma_4}\hat{\sigma_5} & \hat{\sigma_5}^2\end{bmatrix} = \begin{bmatrix}2444 & 1676 & 1502 & 1116 & 1001\\ 1676 & 1965 & 1761 & 1308 & 1173\\ 1502 & 1761 & 2696 & 2003 & 1796\\ 1116 & 1308 & 2003 & 2544 & 2281\\ 1001 & 1173 & 1796 & 2281 & 3494\end{bmatrix}$$

(ix) 이분산 Toeplitz(2)

$$\begin{bmatrix}\hat{\sigma_1}^2 & \hat{p}_1\hat{\sigma_1}\hat{\sigma_2} & 0 & 0 & 0\\ \hat{p}_1\hat{\sigma_1}\hat{\sigma_2} & \hat{\sigma_2}^2 & \hat{p}_1\hat{\sigma_2}\hat{\sigma_3} & 0 & 0\\ 0 & \hat{p}_1\hat{\sigma_2}\hat{\sigma_3} & \hat{\sigma_3}^2 & \hat{p}_1\hat{\sigma_3}\hat{\sigma_4} & 0\\ 0 & 0 & \hat{p}_1\hat{\sigma_3}\hat{\sigma_4} & \hat{\sigma_4}^2 & \hat{p}_1\hat{\sigma_4}\hat{\sigma_5}\\ 0 & 0 & 0 & \hat{p}_1\hat{\sigma_4}\hat{\sigma_5} & \hat{\sigma_5}^2\end{bmatrix} = \begin{bmatrix}2316 & 834 & 0 & 0 & 0\\ 834 & 1595 & 848 & 0 & 0\\ 0 & 848 & 2394 & 919 & 0\\ 0 & 0 & 919 & 1876 & 1055\\ 0 & 0 & 0 & 1055 & 3151\end{bmatrix}$$

(x) 이분산 Toeplitz(2)

$$\begin{bmatrix}\hat{\sigma_1}^2 & \hat{p}_1\hat{\sigma_1}\hat{\sigma_2} & \hat{p}_2\hat{\sigma_1}\hat{\sigma_3} & \hat{p}_3\hat{\sigma_1}\hat{\sigma_3} & \hat{p}_4\hat{\sigma_1}\hat{\sigma_5}\\ \hat{p}_1\hat{\sigma_1}\hat{\sigma_2} & \hat{\sigma_2}^2 & \hat{p}_1\hat{\sigma_2}\hat{\sigma_3} & \hat{p}_2\hat{\sigma_2}\hat{\sigma_4} & \hat{p}_3\hat{\sigma_2}\hat{\sigma_5}\\ \hat{p}_2\hat{\sigma_1}\hat{\sigma_3} & \hat{p}_1\hat{\sigma_2}\hat{\sigma_3} & \hat{\sigma_3}^2 & \hat{p}_1\hat{\sigma_3}\hat{\sigma_4} & \hat{p}_2\hat{\sigma_3}\hat{\sigma_5}\\ \hat{p}_3\hat{\sigma_1}\hat{\sigma_4} & \hat{p}_2\hat{\sigma_2}\hat{\sigma_4} & \hat{p}_1\hat{\sigma_3}\hat{\sigma_4} & \hat{\sigma_4}^2 & \hat{p}_1\hat{\sigma_4}\hat{\sigma_5}\\ \hat{p}_4\hat{\sigma_1}\hat{\sigma_5} & \hat{p}_3\hat{\sigma_2}\hat{\sigma_5} & \hat{p}_2\hat{\sigma_3}\hat{\sigma_5} & \hat{p}_1\hat{\sigma_4}\hat{\sigma_5} & \hat{\sigma_5}^2\end{bmatrix} = \begin{bmatrix}2561 & 1750 & 2016 & 1743 & 1822\\ 1750 & 2052 & 1787 & 1736 & 1801\\ 2016 & 1787 & 2667 & 1959 & 2284\\ 1743 & 1736 & 1959 & 2467 & 2174\\ 1822 & 1801 & 2284 & 2174 & 3286\end{bmatrix}$$

3.2. 공분산모형의 적합성 여부와 유의성검정

공분산행렬 형태의 적합성여부를 판단하기 위하여 개별 공분산행렬에 대한 우도값과 관련있는 $-2log(L)$와 AIC 값을 산출하였다. 표3.3에서 보듯이 등분산인 경우에는 Toeplitz, 이분산인 경우에도 Toeplitz형태의 $-2log(L)$가 가장 작으며, AIC는 가장 큰 값을 보여준다. 등분산인 경우 모수가 각각 2개인 복합대칭성, AR(1), Toeplitz(2) 중에서 복합대칭성 형태가, 이분산인 경우 모수가 각각 6개인 복합대칭성, AR(1), Toeplitz(2) 중에서도 복합대칭성 형태가 $-2log(L)$과 AIC 기준에서 가장 적합한 것으로 나타났다.

그러나 등분산 하에서, 모수의 수가 5개인 Toeplitz 는 다른 형태의 공분산행렬보다 3개 또는 4개가 많은 모수를 가지고 있으며, 이분산 하에서도 모수의 수가 9개인 이분산 Toeplitz가 다른 형태의 공분산행렬보다 3개 또는 4개가 더 많은 모수를 가진다는 점에서 Toeplitz의 공분산 형태가 다른 형태들 보다 더 적합한 것에 대한 유의성검정을 실시할 필요가 있다.

표 3.3 공분산행렬 형태의 우도에 관련된 값

공분산행렬 형태	모수의 수	$-2log(L)$	AIC
(i) 단순형태	1	3094.4	-3096.4
(ii) 복합대칭성	2	2871.9	-2875.9
(iii) AR(1)	2	2901.8	-2905.8
(iv) Toeplitz(2)	2	3002.5	-3006.5
(v) Toeplitz	5	2859.5	-2869.5
(vi) 이분산 단순형태	5	3091.9	-3101.9
(vii) 이분산 복합대칭성	6	2864.2	-2876.2
(viii) 이분산 AR(1)	6	2892.3	-2904.3
(ix) 이분산 Toeplitz(2)	6	2988.7	-3000.7.
(x) 이분산 Toeplitz	9	2852.7	-2870.7

공분산행렬의 모수에 대한 유의성검정은 포함하는 큰 모형의 우도값(L_1)과 포함되는 작은 모형의 우도값(L_2)을 구한 후, $\chi^2 = 2(log(L_i) - log(L_2))$검정을 통해 이루어지며, 이때 자유도는 큰 모형에는 포함되어 있지만 작은 모형에는 포함되어 있지 않는 모수의 갯수이다.

표3.4는 등분산 하에서 포함하는 모형과 포함되는 모형과의 차이에 대한 우도비검정 결과이다. 복합대칭성, AR(1), Toeplitz(2)는 단순형태의 공분산행렬보다 더욱 적합하며, Toeplitz는 복합대칭성, Toeplitz(2) 보다 더욱 적합한 모형으로 나타났다. 따라서 등분산 중에서는 Toeplitz의 공분산 형태가 가장 적합한 모형으로 판명된다.

표 3.4 등분산인 경우 공분산행렬 모수의 차이에 대한 우도비 검정

포함하는 모형	포함되는 모형	자유도	χ^2	유의확률
(ii) 복합대칭성	(i) 단순형태	1	222.5	< 0.0001
(iii) AR(1)	(i) 단순형태	1	192.6	< 0.0001
(iv) Toeplitz(2)	(i) 단순형태	1	91.9	< 0.0001
(v) Toeplitz	(ii) 복합대칭성	3	12.4	0.0061
(v) Toeplitz	(iv) Toeplitz(2)	3	143.0	< 0.0001

표3.5는 이분산 하에서 포함하는 모형과 포함되는 모형과의 차이에 대한 우도비검정 결과이다. 이분산 경우에도 이분산 복합대칭성, 이분산 AR(1), 이분산 Toeplitz(2)는 이분산 단순형태보다 더 적합하며, 이분산 Toeplitz는 이분산 복합대칭성, 이분산 Toeplitz(2)보다 더욱 적합한 모형으로 나타났다. 따라서 이분산 중에서도 이분산 Toeplitz가 가장 적합한 형태인 것을 알 수 있다.

표 3.5 이분산인 경우 공분산행렬 모수의 차이에 대한 우도비 검정

포함하는 모형	포함되는 모형	자유도	χ^2	유의확률
(vii) 이분산 복합대칭성	(vi) 이분산 단순형태	1	227.7	< 0.0001
(viii) 이분산 AR(1)	(vi) 이분산 단순형태	1	199.6	< 0.0001
(ix) 이분산 Toeplitz(2)	(vi) 이분산 단순형태	1	103.2	< 0.0001
(x) 이분산 Toeplitz	(vii) 이분산 복합대칭성	3	11.5	0.0093
(x) 이분산 Toeplitz	(ix) 이분산 Toeplitz(2)	3	136.0	< 0.0001

표3.6은 대응되는 이분산과 등분산의 경우, 공분산행렬 모수의 차이에 대한 우도비검정 결과이다. 이분산 AR(1)과 이분산 Toeplitz(2)는 등분산 AR(1)과 등분산 Toeplitz(2)보다 유의하지만, 이분산 Toeplitz는 등분산 Toeplitz보다 유의하지 않다. 종합적으로 판단해서 등분산 Toeplitz가 이 실험자료에 가장 적합한 공분산행렬 형태로 최종적으로 판단된다.

표 3.6 대응되는 이분산과 등분산의 공분산행렬 모수의 차이에 대한 우도비검정

포함하는 모형	포함되는 모형	자유도	χ^2	유의확률
(vi) 이분산 단순형태	(i) 단순형태	4	2.5	0.6446
(vii) 이분산 복합대칭성	(ii) 복합대칭성	4	7.7	0.1032
(viii) 이분산 AR(1)	(iii) AR(1)	4	9.5	0.0497
(ix) 이분산 Toeplitz(2)	(iv) Toeplitz(2)	4	13.8	0.0080
(x) 이분산 Toeplitz	(v) Toeplitz	4	6.8	0.1468

3.3. 적정 공분산모형 산출

적합성 여부와 유의성검정 결과 등분산 Toeplitz가 최적 공분산행렬 형태로 설정되었으며, 그 형태와 추정된 분산과 시점들 간의 상관계수는 표3.1의 (v) Toeplitz와 같다. 등분산 Toeplitz에서 1시점 차이 간의 상관계수는 0.76으로 어느 정도 높은 관련을 보이고 있지만, 2 시점 간의 상관계수는 0.775로서 1 시점 차이보다 미미하지만 오히려 약간 더 높은 연관성을 보여준다. 3시점 차이는 0.69로 관련 정도가 떨어지며, 4시점 차이는 0.64로 상대적으로 낮은 연관성을 보여준다. 그렇지만 시점들 간의 연관 정도는 0.64 -0.775로 대체적으로 높은 편이다. 즉 각 시점에서 콜레스테롤 수치는 이전 시점의 수치들에 의하여 상당한 영향을 받는다고 할 수 있다.

공분산행렬 형태를 등분산 Toeplitz 로 설정한 모형을 바탕으로 모수인자의 효과에 대한 분산분석 결과는 표3.7과 같다. 처리 및 처리와 시간과의 교호작용 효과는 유의하지 않으나, 시간에 대한 효과는 대단히 유의한 것으로 판명되었다.

참고로 등분산 Toeplitz 이외의 모형들을 가정한 공분산구조 모형에서도 모수인자들의 추정량과 추정량에 대한 유의성검정 결과는 등분산 Topeplitz 의 경우와 거의 차이가 없는 것으로 나타났다. 이는 이 실험자료에서는 공분산구조 모형의 차이가 모수인자의 추정량에는 거의 영향을 주지 않는 것으로 판단된다.

표 3.7 모수인자의 효과에 대한 분산분석표

요인	분자의 자유도	분모의 자유도	F값	유의확률
처리	2	57	0.91	0.4085
시간	4	228	6.74	< 0.0001
처리×시간	8	228	1.44	0.1805

3종류의 식이요법의 차이에 대한 다중비교의 결과는 표3.8에서 알 수 있듯이 3처리간의 차이는 거의 없는 것으로 나타났다. 이는 표3.7의 분산분석표의 처리에 대한 유의확률이 0.41인 것을 봐서도 알 수 있다.

표 3.8 처리 간의 차이에 대한 다중비교

처리간의 차이	추정치	표준오차	t 값	유의확률
I - II	10.85	13.74	0.79	0.4329
I - III	-7.85	14.07	-0.56	0.5790
II - III	-18.70	13.97	-1.34	0.1859

4. 결론

3종류의 식이요법을 복용한 후 5주간에 걸친 콜레스테롤에 관한 반복측정자료를 바탕으로 혼합모형의 분석기법을 적용하여 분석한 결과, 등분산 Toeplitz가 해당자료에 가장 적합한 공분산행렬 형태로 선택되었다. 이 경우 시점들 간의 상관관계는 상당히 높게 나타났으며, 그 중 2시점 간에서 가장 높은 상관관계를 보여주었으며, 다음으로 1시점 간, 3시점 간, 마지막으로 4시점간의 상관관계가 점차 낮음을 보이지만, 시점들 간의 관련 정도는 대체적으로 0.64 - 0.775로 높은 편이다. 공분산행렬 형태를 등분산 Toeplitz 로 채택하여 분석한 결과, 처리 및 처리와 시간과의 교호작용의 효과는 유의하지 않았으나 시간 효과는 대단히 유의한 것으로 판명되었다. 이는 반복측정된 자료들 간에 높은 관련성을 보여주는 것이라고 말할 수 있다.

참고문헌

조진남 (2009). 체중감량자료에 대한 적정 공분산형태모형 산출에 관한 실증연구. <한국데이터정보과학회지>, **20**, 377-385.

조진남, 백재욱 (2009). 변량계수모형의 식이요법 실험자료에 관한 사례연구. <한국데이터정보과학회지>, **20**, 787-796.

Brown, H. and Kempton, R. A. (1994). The Application of REML in clinical trials. *Statistics in Medicine*, **16**, 1601-1617.

Brown, H. and Prescott, R. (1999). *Applied mixed models in medicine*, John Wiley & Sons Inc, New York.

Diggle, P. J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics*, **43**, 1255-1258.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). *Applied longitudinal analysis*, John Wiley & Sons Inc, New York.

Frees. E. D. (2006). *Longitudinal and panel data*, Cambridge University Press, New York.

Hand, D. and Crowder, M. (1996). *Practical longitudinal data analysis,* Chapman & Hall, London.

Hartley, H. O. and Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, **54**, 93-108.

Littell, R. C., Milliken, G. A., Stroup, W. W. and Wolfinger, R. D. (1996). *SAS System for mixed models*, SAS Institute Inc., N.C., U.S.A.

Longford, N. T. (1993). *Random coefficient models*, Oxford University Press, Oxford.

Satterthwaite, F. E (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, **2**, 110-114.

Verbeke, G, and Molenberghs G. (2000). *Linear mixed models for longitudinal data*, Springer Verlag, New York.

Journal of the Korean Data & Information Science Society
2010, **21**(6), 1263–1270

한국데이터정보과학회지

A statistical analysis on the selection of the optimal covariance matrix pattern for the cholesterol data

Jinnam Jo[1] · Jai Wook Baik[2]

[1]Department of Information & Statistics, Dongduk Women's University
[2]Department of Information Statistics, Korea National Open University

Received 2 October 2010, revised 21 November 2010, accepted 23 November 2010

Abstract

Sixty patients were divided into three groups. Each group of twenty persons had fed on different diet foods over 5 weeks. Cholesterol had been measured repeatedly five times at an interval of a week during 5 weeks. It resulted from mixed model analysis of repeated measurements data that homogeneous toeplitz covariance matrix pattern was selected as the optimal covariance pattern. The correlations between measurements of different times for the covariance matrix are somewhat highly correlated as 0.64 -0.78. Based upon the homogeneous toeplitz covariance pattern model, the time effect was found to be highly significant, but the treatment effect and treatment-time interaction effect were found to be insignificant.

Keywords: Covariance matrix pattern, fixed factor, likelihood function, mixed model, repeated measurements data.

[1] Professor, Department of Information & Statistics, Dongduk Women's University, Seoul 136-714, Korea.
[2] Corresponding author: Professor, Department of Information Statistics, Korea National Open University, Seoul 110-791, Korea. E-mail: jbaik@knou.ac.kr