Journal of the Korean Data & Information Science Society
2009, **20**(5), 787–796

한국데이터정보과학회지

변량계수모형의 식이요법 실험자료에 관한 사례연구

조진남[1] · 백재욱[2]

[1] 동덕여자대학교 정보대학 정보통계학과 · [2] 방송통신대학교 정보통계학과

접수 2009년 6월 23일, 수정 2009년 8월 14일, 게재확정 2009년 9월 4일

요 약

이 논문에서는 반복측정치에 대한 분석모형 중, 혼합모형의 일종인 변량계수모형에 대하여 이론적으로 고찰한다. 특히 혼합모형의 설정, 모수 추정에 대하여 통계적으로 고찰하고 변량계수모형에 대한 가능한 모형을 열거하며, 그에 따르는 추정과 검정을 논의한다. 사례연구로 식이요법자료를 대상으로 가능한 변량계수모형을 적용하여 추정 및 검정을 실시한 결과, 고정인자인 사전값, 처리, 키 및 시간들의 인자는 체중감소에 대단히 유의함을 보여주었지만, 나이와 혈압은 유의하지 않았다. 처리효과에 있어서는 식이요법과 운동을 병행했을 때의 처리가 식이요법만 실시했을 때의 처리보다 체중이 더 감소했음을 알 수 있으며, 시간에 따른 체중감소의 효과는 삼차함수의 관계가 성립된다. 변량인자로는 개체효과는 유의하며 개체별 시간에 대한 교호작용의 효과는 차수가 높아질수록 급속도로 감소하여 3차함수 관계가 적절한 모형으로 최종 선택되었다.

주요용어: 반복측정 데이터, 변량계수모형, 변량인자, 우도함수, 혼합모형.

1. 서론

의학, 약학 및 보건학 등의 임상실험에서 개체들을 대상으로 반복적으로 실험을 실시한 후 처리효과를 규명하고자 할 때 시간이 중요한 변수가 된다. 실험을 실시할 때 시점들이 일정하게 정해져 있는 경우와 일정하게 고정되어 있지 않을 경우의 임의의 시점으로 구분할 수 있다. 반복처리 시점이 고정된 경우 반복측정자료의 통계적 기법을 이용하여 실험 자료를 분석하는데, 이는 시점들 간의 상관관계를 규정짓는 공분산 형태 (covariance pattern)에 따라 분석기법을 달리한다. 측정시점이 고정된 반복측정자료를 이용하여 적정 공분산형태를 찾는 실증연구로는 조진남 (2009)이 있다. 반면에, 임의의 시점들인 경우 변량계수모형 (random coefficient model)을 적용하여 실험자료를 분석하여 결과를 도출한다. 변량계수모형에서는 실험개체들의 처리에 대한 시간에 따른 반응값이 어떤 패턴을 보이는지 규명하는 것이 중요하다. 변량계수모형을 이용하여 반복측정자료를 분석한 문헌들은 Brown과 Komtom (1999), Choi (2008), Frees (2006), Hay 등 (1998), Longford (1993), Smyth 등 (1997) 등이 있다. 변량계수모형은 고정인자 (fixed factor)와 변량인자 (random factor)를 함께 포함하므로 혼합모형 (mixed model)에 속한다. 따라서 이 논문에서는 일반적 혼합모형을 설정하고 그에 따르는 통계적 성질을 간략하게 살펴보고, 다음으로 변량계수모형을 설정하고 해당 모형에 대한 추정 및 검정의 이론적 성질을 규명한다. 마지막으로 사례분석으로 식이요법자료를 이용하여 가능한 변량계수모형을 설정하며, 개별 모형에 대한 유의성 검정을 실시하며, 적합성 판단기준에 따라 최적모형을 산출한다. 도출된 최적모형을 바탕으로 반응값과 시간과의 연관성을 규명해보고자 한다.

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

2. 이론적 배경

2.1. 혼합모형의 모수 추정

일반적으로 혼합모형은 다음과 같다.

$$y_i = \mu + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_p x_{ip} + \beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_q z_{iq} + e_i, \tag{2.1}$$

$$i = 1, 2, \cdots, n.$$

n 개의 관찰치가 주어졌을 때 행렬형태로 표시하면 다음과 같다.

$$\underline{y} = X\underline{\alpha} + Z\underline{\beta} + \underline{e}. \tag{2.2}$$

여기서 $\underline{y} = (y_1, y_2, \cdots, y_n)'$는 $n \times 1$의 관찰치 벡터, X는 $n \times (p+1)$크기의 고정인자들의 디자인 행렬 (design matrix), $\underline{\alpha} = (\mu, \alpha_1, \alpha_2, \cdots, \alpha_p)'$는 고정인자 계수 벡터, Z는 $n \times q$크기의 변량인자 들의 디자인 행렬, $\underline{\beta} = (\beta_1, \beta_2, \cdots, \beta_q)'$는 평균 $\underline{0}$, $q \times q$크기의 분산공분산행렬 G를 가지는 다변량 정규분포를 하는 변량인자 계수 벡터, 그리고 $\underline{e}$는 평균 $\underline{0}$, 분산공분산행렬 R인 다변량 정규분포를 하는 $n \times 1$ 크기의 오차벡터이다. 따라서 이 모형에서 $\underline{y}$의 분산공분산행렬 V는 다음과 같다.

$$V = var\left(\underline{y}\right) = ZGZ' + R. \tag{2.3}$$

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

$$L = (2\pi)^{-\left(1/2\right)n} |V|^{-\left(1/2\right)} \exp\left[-\frac{1}{2}\left(\underline{y} - X\underline{\alpha}\right)' V^{-1} \left(\underline{y} - X\underline{\alpha}\right)\right], \tag{2.4}$$

$$\log(L) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\left[\log|V| + \left(\underline{y} - X\underline{\alpha}\right)' V^{-1}\left(\underline{y} - X\underline{\alpha}\right)\right]. \tag{2.5}$$

이 우도함수를 최대화 시키는 제한최우추정법 (REML: restricted maximum likelihood), 또는 반복일반화최소자승법 (IGLS: iterative generalized least squares)에 의하여 고정인자효과와 변량인자 효과의 추정량과 추정량의 분산은 다음과 같다.

$$\underline{\hat{\alpha}} = \left(X'V^{-1}X\right)^{-1} X'V^{-1}\underline{y}. \tag{2.6}$$

$$var\left(\underline{\hat{\alpha}}\right) = \left(X'V^{-1}X\right)^{-1}, \tag{2.7}$$

$$\underline{\hat{\beta}} = \left(Z'R^{-1}Z + G^{-1}\right)^{-1} ZR^{-1}\left(\underline{y} - X\underline{\alpha}\right). \tag{2.8}$$

$$var\left(\underline{\hat{\beta}}\right) = GZ'\left[V^{-1} - V^{-1}X\left(X'V^{-1}X\right)^{-1}X'V^{-1}\right]ZG. \tag{2.9}$$

2.2. 변량계수 모형의 설정

다음은 여러 가지 가능한 경우의 변량계수모형들이다.

1) 개체효과를 고려하지 않을 경우

$$y_{ij} = \mu + b\,(pre)_i + t_k + m\,(time)_{ij}\,. + e_{ij} \tag{2.10}$$

여기서 y_{ij}는 i번째 개체의 j번째 시점에서의 반응값, $(pre)_i$는 i번째 개체의 처리 전 반응값이며, b는 그에 대한 회귀계수이다, t_k는 k번째 처리효과, $(time)_{ij}$는 i번째 개체의j번째 시점을 의미하며, m은 $(time)_{ij}$에 대한 회귀계수이다. 여기서 회귀계수 b, t_k, m은 고정인자 효과 들이다. 따라서 이 모형은 고정인자효과 모형 (fixed effect model)에서 추정 및 검정을 할 수 있다. 모형1은 개별 개체의 효과를 고려하지 않고 전체적으로 시간에 대한 반응값의 변화를 알고자 할 때 이용한다.

2) 선형관계인 경우

$$y_{ij} = \mu + b\,(pre)_i + t_k + p_i + m\,(time)_{ij} + (pm)_i\,(time)_{ij} + e_{ij}. \tag{2.11}$$

여기서 p_i는 i번째 개체효과, $(pm)_i$는 i번째 개체의 시간에 대한 기울기값을 의미하며, p_i와 $(pm)_i$는 변량인자효과 (randon factor effect)의 계수로 서로 독립이 아니며 이변량 정규분포를 따른다. 즉

$$\begin{pmatrix} p_i \\ pm_i \end{pmatrix} \sim N\,(0, G)$$

이며, 공분산행렬 G는 다음과 같다.

$$G = \begin{bmatrix} \sigma_p^2 & \sigma_{p,pm} \\ \sigma_{p,pm} & \sigma_{pm}^2 \end{bmatrix}. \tag{2.12}$$

여기서 σ_p^2, σ_{pm}^2은 각각 p_i, $(pm)_i$의 분산성분이며, $\sigma_{p,pm}$은 p_i, $(pm)_i$의 공분산이다. 분산성분의 추정값이 0 보다 클때는 의미가 있지만, 음수로 추정될 때는 해당인자의 효과는 의미가 없는 것으로 간주한다. b, t_k, m은 고정인자 효과들이지만, p_i, $(pm)_i$은 변량인자효과이므로, 혼합모형을 이용하여 분석한다. 모형2는 각 개체의 효과를 고려한 후, 시간에 따른 반응값의 선형관계를 파악하고자 할 때 이용된다.

3) 키 등 고정인자 효과를 추가할 경우

$$y_{ij} = \mu + b\,(pre)_i + c\,(Height)_i + t_k + p_i + m\,(time)_{ij} + (pm)_i\,(time)_{ij} + e_{ij}. \tag{2.13}$$

키, 혈압, 나이 등 다른 고정인자들이 반응값에 영향을 미칠 때 사용한다. 따라서 b, c, t_k, m은 고정인자효과, p_i, $(pm)_i$는 변량인자효과이며, 시간과 반응값의 관계는 선형관계이다.

4) 처리와 시간간의 교호작용 추가

$$y_{ij} = \mu + b\,(pre)_i + t_k + p_i + m\,(time)_{ij} + (tm)_k\,(time)_{kj} + (pm)_i\,(time)_{ij} + e_{ij}. \tag{2.14}$$

처리와 시간 간의 교호작용인자 $(time)_{kj}$를 추가한 모형이다. 이 모형에서도 b, t_k, m, $(tm)_k$ 는 고정인자효과이며, p_i, $(pm)_i$는 변량인자효과이며, 시간과 반응값은 선형관계이다.

5) 교호작용 없는 이차함수관계

$$y_{ij} = \mu + b\,(pre)_i + t_k + p_i + m_l\,(time)_{ij} + m_q\,(time)^2 + (pm_l)_i\,(time)_{ij} + (pm_q)_i\,(time)_{ij}^2 + e_{ij}. \tag{2.15}$$

반응값이 개체 별로 시간과의 관계가 1차 및 2차 함수와의 관계를 설정한 모형이며, 처리와 시간과의 교호작용은 고려하지 않는다. 여기서 고정인자효과는 b, t_k, m_l, m_q이며, 변량인자효과 p_i, $(pm_l)_i$, $(pm_q)_i$는 다변량정규분포를 한다. 구체적으로

$$\begin{pmatrix} p_i \\ (pm_l)_i \\ (pm_q)_i \end{pmatrix} \sim N\,(0, G)\,,$$

이며, 공분산행렬 G 는 다음과 같이 구성된다.

$$G = \begin{bmatrix} \sigma_p^2 & \sigma_{p,pm_l} & \sigma_{p,pm_q} \\ \sigma_{p,pm_l} & \sigma_{pm_l}^2 & \sigma_{pm_l,pm_q} \\ \sigma_{p,pm_q} & \sigma_{pm_l,pm_q} & \sigma_{pm_q}^2 \end{bmatrix}. \tag{2.16}$$

$\sigma_p^2, \sigma_{pm_l}^2, \sigma_{pm_q}^2$는 각각 p_i, $(pm)_i$, $(pm_q)_i$의 분산성분이며, $\sigma_{p,pm}\sigma_{p,pm_q}\sigma_{pm_l,pm_q}$는 공분산성분들이다. 이 경우 2차항에 해당하는 분산성분 $\sigma_{pm_q}^2$이 음수로 추정될 경우 해당 변수는 모형에서 제거되며, 더 이상 고차항의 관계는 존재하지 않는다.

6) 교호작용 없는 삼차함수관계

$$\begin{aligned} y_{ij} = \mu + b\,(pre)_i + t_k + p_i + m_l\,(time)_{ij} + m_q\,(time)_{ij}^2 + m_c\,(time)_{ij}^3 \\ + (pm_l)_i\,(time)_{ij} + (pm_q)_i\,(time)_{ij}^2 + (pm_c)_i\,(time)_{ij}^3 + e_{ij}. \end{aligned} \tag{2.17}$$

반응값이 개체별 시간과의 관계가 1차, 2차 및 3차 함수와의 관계를 설정한 모형이며, 모형 5와과 같이 처리와 시간과의 교호작용은 고려하지 않는다. 고정인자효과는 b, t_k, m_l, m_q, m_c이며, 변량인자효과 p_i, $(pm_l)_i$, $(pm_q)_i$, $(pm_c)_i$는 다음과 같이 가정한다. 즉

$$\begin{pmatrix} p_i \\ (pm_l)_i \\ (pm_q)_i \\ (pm_c)_i \end{pmatrix} \sim N(0, G).$$

여기서

$$G = \begin{bmatrix} \sigma_p^2 & \sigma_{p,pm_l} & \sigma_{p,pm_q} & \sigma_{p,pm_c} \\ \sigma_{p,pm_l} & \sigma_{pm_l}^2 & \sigma_{pm_l,pm_q} & \sigma_{pm_l,pm_c} \\ \sigma_{p,pm_q} & \sigma_{pm_l,pm_q} & \sigma_{pm_q}^2 & \sigma_{pm_q,pm_c} \\ \sigma_{p,pm_c} & \sigma_{pm_l,pm_c} & \sigma_{pm_q,pm_c} & \sigma_{pm_c}^2 \end{bmatrix} \tag{2.18}$$

이며, 각 분산 및 공분산 성분은 2차함수 모형의 (2.17)과 같으며 $(pm_c)_i$의 분산성분 $\sigma_{pm_c}^2$과 그에 해당하는 공분산성분이 추가된다. 이 경우 3차항에 해당하는 분산성분 $\sigma_{pm_c}^2$이 음수로 추정될 경우 해당 변수는 모형에서 제거된다.

2.3. 변량계수모형의 통계적 추론

변량계수모형은 혼합모형에 속하므로 추정량은 혼합모형의 추정량을 구하는 과정과 같은 방법으로 구한다. 이 때 오차항의 분산 $R = \sigma^2 I$으로 가정하므로 변량인자 계수 $\widehat{\underline{\beta}}$은 식 (2.6)로부터 다음과 같이 도출된다.

$$\widehat{\underline{\beta}} = \left(Z'Z + G^{-1}/\sigma^2\right)^{-1} Z\left(\underline{y} - X\underline{\alpha}\right). \tag{2.19}$$

이때 우도함수를 분산성분에 관하여 미분한 결과는 비선형이므로, 분산성분의 추정은 뉴턴-라프슨 알고리즘 (Newton-Raphson algorithm)을 이용하여 얻은 Fisher의 정보행렬 (information matrix)에서 얻을 수 있다.

사전값, 키, 처리, 시간 등 고정인자 효과의 검정은 분산분석표에서 Satterthwaite의 근사 자유도 (approximate degree of freedom)를 근거로 하여 각 회귀계수를 검정하여 유의성을 판정한다.

일반적으로 각 개체의 관찰치 수가 극히 적을 때, 고정효과모형을 이용하여 분석할 경우, 개체별로 시간에 대한 기울기값 들은 심히 왜곡된 결과를 초래할 수 있지만, 변량계수모형을 이용하여 분석할 경우, 개체효과 p_i와 기울기에 해당하는 개체와 시간과의 교호작용의 효과 $(pm)_i$의 추정치 자체는 중요하지 않으므로 극단치 값을 방지해 주는 이점이 있다.

3. 사례연구

사례연구로 식이요법에 관련된 실험자료에 변량계수모형을 적용하여 최적의 변량계수모형을 산출하고 해당모형을 바탕으로 고정인자 효과와 변량인자 효과를 도출하며, 시간에 따른 반응치의 값에 대한 적절한 패턴을 살펴보고자 한다.

식이요법과 운동이 체중감량에 미치는 효과에 관련된 논문은 장은재 등 (1997)이 있으며, 구체적인 실험내용은 다음과 같다. 25명의 여대생을 대상으로 체중감량에 대한 식이요법과 운동을 병행하는 처리 (이하 처리 A)와 식이요법만 실시하는 처리 (이하 처리 B)에 대한 효과를 비교하고자 무작위로 2개의 처리그룹으로 나누어서 8주간에 걸쳐서 반복측정실험을 실시하였다. 반응변수는 실험실시 후 1주, 6주째를 제외한 매주 반복 측정된 체중값이며, 고정인자로는 사전값, 처리, 시간, 나이, 키, 혈압이 해당되며, 개체와 개체별 시간효과는 변량인자에 속한다. 그러나 이 실험에서 측정하지 못한 4주, 5주, 7주째의 결측치 들이 존재하는데, 결측치 여부에 관계없이 회귀모형을 기반으로 한 혼합모형을 사용하여 추정치의 결과를 산출한다. 실제 데이터분석은 혼합모형에 관련된 통계프로그램인 SAS의 PROC MIXED를 이용하였으며, 실험데이터는 부록에 수록하였다.

3.1. 분석결과

표 3.1은 변량계수모형에서 모형 1 – 모형 4까지의 추정결과를 보여준다. 모형1 은 고정인자들인 사전 체중값, 2개의 처리간의 차이 및 시간의 흐름이 체중변화에 유의적으로 판명되었다. 모형2는 모형1의 고정인자와 변량인자인 개체효과 p_i와 개체별 시간효과 $(pm)_i$로 구성되는데, 변량인자의 분산공분산행렬 G의 분산값이 양수이므로 p_i와 $(pm)_i$는 체중감량에 의미가 있으므로, 모형에 포함시키는 것이 적절하다. 또한 모형2의 오차항의 분산이 모형1의 분산값보다 훨씬 작으므로, 모형2가 모형1보다 더 적절한 모형이라고 판단된다. 모형3에서 나이와 혈압은 체중감량과는 관련이 없지만, 키는 유의한 것으로 나타났으므로 키만 추가한 모형 3-1이 적절한 모형이 된다. 모형4에서 시간과 처리와의 교호작용 효과는 유의하지 않다. 따라서 표 3.1에 있는 모형들 중에서 모형 3-1이 적절한 모형으로 판단된다.

다음으로 모형 3-1를 기준으로 시간과 체중간의 다항관계를 파악하고자 한다. 표 3.2에서 모형 3-1은 선형관계, 모형 5는 처리와 시간과의 교호작용이 없는 2차 함수관계, 모형 6은 처리와 시간과의 교호작용이 없는 3차 함수관계로 설정한 모형들의 추정 결과이다. 추정결과 고정인자의 시간에 대한 2차함수와 3차함수는 유의적으로 나타났으며, 변량인자 p_i, $(pm_l)_i$, $(pm_q)_i$, $(pm_c)_i$ 의 분산값들도 G 행렬의 대각선상에 나타난 바와 같이 양수로 판명되었으므로, 이 인자들은 의미가 있다고 간주한다. 그러나 분산의 크기는 차수가 거듭될수록 적으므로 고차항일수록 영향력은 미미해진다. 마지막으로 시간에 대한 4차항의 관계를 설정하여 분석한 결과 추정치가 수렴되지 않으므로 시간에 대한 4차항의 모형은 적절하지 않다. 따라서 이 사례연구의 식이요법의 자료에서는 처리와 시간과의 교호작용이 없는 3차 함수관계인 모형6을 최종모형으로 설정한다.

모형적합성 판정에서 기준으로 우도값에 관련된 $-2\log(L)$, AIC 와 오차분산을 채택하였다. $-2\log(L)$ 과 오차분산은 적을수록 좋은 모형이며, AIC는 클수록 좋은 모형이다. 각 모형별로 $-2\log(L)$, AIC와 오차분산 값을 표 3.3에 적시하였다. 이 기준에 따라 고정인자만 포함시킨 모형1보다 변량인자의 개체효과 및 개체와 시간과의 교호작용 효과들을 포함시킨 혼합모형에서 $-2\log(L)$, AIC, 오차분산

표 3.1 모형 1 - 모형 4 의 추정결과

모형	고정인자	회귀계수 (표준오차)	유의확률	G 행렬과 오차분산값
모형1	절편	2.689(1.470)	0.070	
	사전값	0.952(0.022)	< 0.001	
	처리	-1.063(0.322)	0.001	3.041
	시간	-0.204(0.073)	0.007	
모형2	절편	-2.112(1.732)	0.236	
	사전값	1.022(0.027)	< 0.001	$\begin{bmatrix} 0.517 & 0.019 \\ 0.019 & 0.093 \end{bmatrix}$
	처리	-0.578(0.385)	0.147	
	시간	-0.208(0.066)	0.004	0.387
모형3	절편	11.180(5.951)	0.076	
	사전값	1.069(0.031)	< 0.001	$\begin{bmatrix} 0.352 & 0.040 \\ 0.040 & 0.093 \end{bmatrix}$
	처리	-0.803(0.373)	0.044	
	나이	-0.027(0.082)	0.740	0.388
	키	-0.084(0.014)	0.048	
	혈압	-0.020(0.014)	0.166	
	시간	-0.208(0.066)	0.004	
모형 3-1	절편	10.382(5.628)	0.079	
	사전값	1.052(0.028)	< 0.001	$\begin{bmatrix} 0.382 & 0.008 \\ 0.008 & 0.093 \end{bmatrix}$
	처리	-0.798(0.363)	0.040	
	키	-0.090(0.039)	0.033	0.388
	시간	-0.208(0.066)	0.004	
모형4	절편	10.323(5.629)	0.081	
	사전값	1.052(0.028)	< 0.001	$\begin{bmatrix} 0.382 & 0.008 \\ 0.008 & 0.093 \end{bmatrix}$
	처리	-0.713(0.372)	0.069	
	키	-0.090(0.039)	0.033	0.388
	시간	-0.130(0.099)	0.205	
	시간*처리	-0.139(0.133)	0.305	

표 3.2 시간에 대한 다항관계 모형의 추정결과

모형	고정 인자	회귀계수 (표준오차)	유의 확률	G 행렬과 오차분산값
모형 3-1	절편	10.382(5.628)	0.079	
	사전값	1.052(0.028)	< 0.001	$\begin{bmatrix} 0.382 & 0.008 \\ 0.008 & 0.093 \end{bmatrix}$
	처리	-0.798(0.363)	0.040	
	키	-0.090(0.039)	0.033	0.388
	시간	-0.208(0.066)	0.004	
모형 5	절편	11.398(5.621)	0.056	
	사전값	1.053(0.028)	< 0.001	$\begin{bmatrix} 1.332 & -0.537 & 0.036 \\ -0.537 & 0.401 & -0.022 \\ 0.036 & -0.022 & 0.001 \end{bmatrix}$
	처리	-0.802(0.362)	0.038	
	키	-0.092(0.039)	0.029	
	시간	-0.595(0.194)	0.005	0.338
	(시간)^2	-0.038(0.016)	0.027	
모형 6	절편	15.379(5.644)	0.013	
	사전값	1.050(0.028)	< 0.001	$\begin{bmatrix} 3.679 & -2.094 & 0.356 & -0.018 \\ -2.094 & 1.526 & -0.267 & 0.014 \\ 0.356 & -0.267 & 0.057 & -0.003 \\ -0.018 & 0.014 & -0.003 & 0.0002 \end{bmatrix}$
	처리	-0.735(0.360)	0.054	
	키	-0.091(0.039)	0.029	
	시간	-3.673(0.587)	< 0.001	
	(시간)^2	0.748(0.130)	< 0.001	0.212
	(시간)^3	-0.049(0.009)	< 0.001	

값이 훨씬 개선되었으며, 기타 혼합모형들 중에서 기준값의 비교에서도 모형6이 가장 좋은 모형으로 판명되었으므로 최종적으로 모형6을 적정모형으로 선택한다.

표 3.3 모형 적합성에 관련된 기준값 결과

모형	$-2\log(L)$	AIC	오차분산값
모형1	555.8	-557.8	3.041
모형2	373	-381	0.387
모형3	380.7	-388.7	0.388
모형3-1	372.8	-380.8	0.388
모형4	373.9	-381.9	0.388
모형 5	370.2	-384.2	0.338
모형 6	348.4	-370.4	0.212

3.2. 최적모형에 대한 통계적 분석

모형6에 대한 고정인자의 분산분석표는 표 3.4와 같다. 즉 사전값, 처리, 키 및 시간인자는 대단히 유의한 것으로 나타났다.

표 3.4 모수인자의 효과에 대한 분산분석표

요인	분자의 자유도	분모의 자유도	F 값	유의확률
사전값	1	20.8	1421.97	< 0.001
처리	1	20.9	4.17	0.054
키	1	20.7	5.48	0.029
시간	1	22.2	39.13	< 0.001
$(시간)^2$	1	21.2	33.22	< 0.001
$(시간)^3$	1	20.7	31.03	< 0.001

표 3.5에서 알 수 있듯이 처리간의 차이는 대단히 유의하며, 식이요법과 운동을 병행했을 때 (처리 A)의 체중이 식이요법만 실시했을 때 (처리 B)의 체중보다 0.735 Kg 감소했음을 알 수 있다.

표 3.5 두 처리간의 차이에 대한 검정

처리의 차이	추정치	표준오차	t 값	유의확률
A-B	-0.735	0.360	-2.04	0.054

시간에 따른 체중감소의 효과는 표 3.6에서 보듯이 삼차함수의 관계가 성립된다. 즉, 처리 후 4주까지는 점차 선형 관계로 감소하다가 5주에서 7주째까지는 다소 증가하였으며, 8주째에는 다시 감소하는 경향을 보였다. 마지막으로 변량인자효과 p_i, $(pm_l)_i$, $(pm_q)_i$, $(pm_c)_i$의 분산공분산행렬 G의 추정값은

표 3.6 시간에 따른 평균값

시간	처리 A	처리 B	전체 평균
사전값	65.21	64.00	64.68
2 주	63.85	63.32	63.62
3 주	63.00	62.70	62.87
4 주	62.45	62.60	62.51
5 주	63.59	64.00	63.78
7 주	63.32	64.22	63.60
8 주	61.76	62.21	61.96

다음과 같다.

$$G = \begin{bmatrix} 3.679 & -2.094 & 0.356 & -0.018 \\ -2.094 & 1.526 & -0.267 & 0.014 \\ 0.356 & -0.267 & 0.057 & -0.003 \\ -0.018 & 0.014 & -0.003 & 0.0002 \end{bmatrix}$$

4개 변량인자의 분산값은 모두 0보다 큰 양수이므로 의미가 있지만, 개체별 시간에대한 교호작용의 효과는 차수가 높아질수록 분산값이 1.526, 0.057, 0.0002로 급속도로 감소되며, 3차항의 $(pm_c)_i$의 분산값은 0.0002로 거의 0에 가깝게 되므로 효과는 아주 미미하다.

4. 결론

식이요법에 관련된 실험자료를 바탕으로 변량계수모형을 적용하여 분석한 결과, 반응값이 시간과의 관계가 3차 함수가 적절한 형태로 선택되었으며, 개체별 시간이 3차 함수와의 관계가 설정된 모형 6이 최종모형으로 선택되었다. 이 모형에서 고정인자인 사전값, 처리, 키 및 시간들의 인자는 체중감소에 대단히 유의함을 알 수 있지만, 나이와 혈압은 유의하지 않았다. 처리효과에 있어서는 식이요법과 운동을 병행했을 때의 처리가 식이요법만 실시했을 때의 처리보다 체중이 0.798Kg 더 감소했음을 알 수 있으며, 시간에 따른 체중감소의 효과는 삼차함수의 관계가 성립된다. 변량인자인 개체 및 개체와 시간과의 교호작용에서 1차, 2차 및 3차 함수의 관계는 유의하지만, 4차 함수의 관계는 유의하지 않았으며, 개체와 시간과의 교호작용의 효과에서 차수가 높아질수록 분산값이 작아지며, 그 효과는 급속도로 감소함을 보였다.

부록: 식이요법 데이터

개체	처리	나이	키	혈압	사전값	y2	y3	y4	y5	y7	y8
1	A	17	165.5	130	64.9	62.8	62.6	63.1	63.3	64.3	63.0
2	A	22	154.8	110	59.1	57.0	56.0	55.3	56.1	56.3	54.9
3	A	21	158.1	100	62.7	60.9	60.3	59.1	58.6	58.2	57.5
4	A	26	161.6	120	61.0	58.8	58.8	58.1	59.3	59.3	58.2
5	A	20	158.0	110	57.5	56.9	55.8	55.1			53.6
6	A	20	168.3	140	75.2	73.2	71.8	70.3	70.3	71.0	69.9
7	A	20	163.2	110	72.1	69.9	67.8	65.5	64.3	63.0	61.2
8	A	26	153.5	110	57.5	55.8	54.7	55.3	56.1	55.9	55.2
9	A	21	154.1	140	69.3	67.9	66.5	67.0	67.1	.	66.8
10	A	18	152.2	120	72.7	72.7	70.2	70.6	70.8	69.4	69.2
11	A	20	161.1	110	70.1	68.1	68.2	66.6		65.8	66.4
12	A	22	148.6	90	51.3	51.2	50.5	50.9	51.6	52.1	52.1
13	A	21	157.7	130	80.2	80.7	81.3	81.3	82.0	81.2	80.5
14	A	18	154.9	120	59.4	58.0	57.5	56.1			56.2
15	B	18	159.5	130	63.9	61.9	61.0	60.6	51.2	60.1	58.5
16	B	21	166.0	120	65.4	63.4	63.1	63.2	62.7	63.0	62.9
17	B	21	156.8	140	55.0	56.7	56.5	54.7		58.7	57.9
18	B	20	155.5	120	56.3	54.5	55.1	55.3	54.9	.	53.9
19	B	19	160.6	140	73.7	72.4	71.5	70.6	71.3	.	70.7
20	B	19	157.5	90	60.7	61.4	60.4	60.3	60.7	.	60.5
21	B	20	154.1	110	66.4	66.4	65.5	65.6	65.9	.	64.5
22	B	20	171.4	110	72.7	73.1	71.4	71.5	71.7	71.4	71.9
23	B	18	157.6	110	58.3	57	56.0	56.4			56.3
24	B	24	161.7	130	62.5	61.1	61.4	.	60.3	.	59.5
25	B	22	157.5	120	69.1	68.6	67.8	67.8	67.3	67.9	67.7

참고문헌

장은재, 임경아, 한용봉 (1999). 영양교육이 체중조절프로그램에 미치는 효과에 관한연구. <한국식품영양학회지>, **12**, 177-183.

장은재, 조진남, 황종현 (1997). 한국 여대생의 체지방 측정을 통한 측정기기들 간의 비교 연구. <한국식품영양과학회지>, **26**, 514-520.

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

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.

Choi, J. (2008). A marginal logit mixed-effects model for repeated binary response data. *Journal of the Korean Data & Information Science Society*, **19**, 413-420.

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.

Hay, C. R. M., Ludlam, C. A., Lowe, G. D. O., Mayne, E. E., Lee, R. J., Prescott, R. J. and Lee, C. A. (1998). The effect of monoclonal or ion-exchange purified factor VIII concentrate on HIV disease progression: A prospective cohort comparison. *British Journal of Haematology*, **101**, 632-637.

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.

Rao, C. R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, **67**, 112-117.

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

Smyth, J. F., Brown, A., Perren, T., Wilkinson, P., Prescott, R. J., Quinn, K. J. and Tedeschi, M. (1997). Glutathione reduces the toxicity and improves quality of the life of women dianosed with ovarian cancer treated with cisplatin: Results of a double blind, randomized trial. *Annals of Oncology*, **8**, 569-5873.

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

Journal of the Korean Data & Information Science Society
2009, **20**(5), 787–796

한국데이터정보과학회지

A case study on the random coefficient model for diet experimental 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 23 June 2009, revised 14 August 2009, accepted 4 September 2009

Abstract

A random coefficient model is applied when times of the repeated measurements are not fixed in experiments with respect to the subjects. The procedures of the inference of a random coefficient model are same as those of a mixed model. Diet experimental data was used for applying the random coefficient model. Various random coefficient models are investigated for the experimental data, and are compared each other. Finally, optimal random coefficient model would be selected. It resulted from the analysis that for the fixed effect factor, the baseline, treatment, height, and time effect were very significant. The treatment effect of the diet foods and exercises were more effective in losing weight than the effect of the diet foods only. The fixed cubic time effect was very significant. The variance components corresponding to the subject effect, linear time effect, quadratic time effect, and cubic time effect of the random coefficients are all positive. When quartic time effect was added as random coefficients the model did not converge. Thus random coefficients up to the cubic terms was considered as the optimal model.

Keywords: Likelihood function, mixed model, random coefficient model, random factor, repeated measures data.

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