The Korean Communications
in Statistics Vol. 7, No. 3, 2000
pp. 677-686

Improved Algorithm for Case-Deletion Diagnostic in Mixed Linear Models[1)]

Jang-Taek Lee[2)]

Abstract

Outliers may occur with respect to any of the random components in mixed linear models. We develop a use of simple, inexpensive updating formulas to consider the effect of case-deletion for mixed linear models. The method described here requires inversions of an n×n matrix, where n is the number of nonempty cells. A numerical example illustrates the use of computational formulas.

Keyword : Case-deletion; Mixed linear model.

1. 서론

혼합모형은 표본계획의 설정, 제품의 성능을 평가하는 품질관리, 실험연구, 통계적 유전학 등에서 널리 사용되고 있다. 하지만 혼합모형을 현명하게 사용하기 위해서는 혼합모형에 사용된 가정들이 분석자의 요구에 과연 타당한지를 진단하여 볼 필요가 있다. 그리고 혼합모형의 분석은 주로 선형, 이차, 우도방법을 이용하는데, 이 방법들은 모두 이상점에 매우 민감한 것으로 알려져 있다.

이 경우 이상점이란 조사의 대상이 되는 모집단에 속하지 않는다고 의심이 될 정도로 정상범위 밖으로 아주 동떨어진 관찰점을 의미하는데, 보통 데이터를 분석하는 사람이 생각하기에 대부분의 관찰점과 아주 동떨어진 관찰점이거나 가정된 확률분포에서 나올 가능성이 없는 관찰점을 뜻한다.

일반적으로 이상점이 존재하는 이유는 모형이 잘못되면 이상점이 여러 개 발생할 수 있는 잘못된 모형의 선택, 어떤 관찰점이 모형전체에는 큰 영향을 미치지 않고 그 관찰점 자체만이 문제가 되는 모형의 국부적 약점, 데이터셋이 나타내는 현상자체가 가지고 있는 변동성이 지나치게 큰 자연적 변화 등이 주원인이 될 수 있다(최병선, 1997). 따라서 개개의 관측값이 추정모형에 어떠한 영향을 미치고 있느냐를 파악하는 것은 매우 중요한 통계분석의 절차라고 할 수 있으며, 이상점들을 다루는 객관적인 방법들을 이용하여 이상점을 살펴본 후 사용자들은 이상점을 제거하거나 더 많은 자료를 수집하거나 모형의 부적당함을 알 수 있다.

과거 20년간은 확실히 통계모형이 가지고 있는 가정의 타당성 여부를 살펴보는 데 지대한 관심을 가졌던 기간이었다. 이와 같은 움직임의 선구자는 회귀분석에서 케이스 제거진단으로 오늘날에도 널리 사용되는 Cook(1977)이라고 할 수 있으며, 회귀분석의 경우 대부분의 통계패키지는 모두

1) This research was supported by a Korea Science and Engineering Foundation Grant in 1998. (Project No: 981-0105-031-1)

2) Professor, Division of Natural Science, Dankook University, Seoul, 140-714, Korea.
E-mail : jangtaek@hanmail.net

회귀진단에 사용되는 여러 가지 통계량들을 제공하고 있다. Cook과 Weisberg(1982), Belsley, Kuh 와 Welsch(1982)의 책은 회귀진단에 대한 이론 및 적용을 상세하게 다루는 대표적인 책이다.

1980년대 후반부터는 본격적으로 회귀모형에서 혼합모형으로 일반화되어서 진단의 문제가 언급되기 시작했는데, Fellner(1986)는 혼합모형에 있어서 분산성분을 추정하는데 이상점의 영향력을 제한최우추정량(REMLE)을 이용하여 구하였다. 또한 Hocking과 Bremer(1987)는 각 셀 당 반복수가 같은 균형혼합모형에서 분산성분의 추정량(AVE 추정량)을 제안하였고, 이 추정량들을 이용한 진단방법을 제안하였다. 그들은 AVE 추정량의 구체적인 모양과 쉽게 계산할 수 있는 방법, AVE 추정량에 포함된 진단정보들을 설명하였는데, 제안된 방법의 단점은 셀 도수가 0인 경우를 점검하지 못하였고, 일반적인 불균형혼합모형에 있어서 AVE 추정량의 통계적 최적성을 논의하지 못했다.

1990년대에는 보다 주목할 만한 논문들이 발표되었는데, Christensen, Pearson과 Johnson(1992)은 REMLE를 이용하여 혼합모형의 케이스 제거 진단 문제를 다루었다. 그들은 케이스 제거진단을 위한 편리한 계산방법 및 고정효과와 분산성분에 대한 진단도구들을 제안하였다. 그리고 Hurtado(1993)는 최우추정량(MLE)을 이용한 혼합모형의 진단에 사용되는 통계량을 제안하였는데, 그의 방법은 분산행렬의 구조가 알려진 경우에 있어서 분산성분의 비율을 관찰자가 아는 경우에는 제안된 통계량을 이용하여 정확하게 한 개의 관측치가 영향력관찰점인가를 알 수 있으며, 분산행렬을 알지 못하는 경우에 몇 가지 가정아래에서 근사적으로 적용 가능하다. 하지만 여러 개의 관측치가 결합영향력 관찰점 인지는 그의 통계량으로도 알 수 없다.

본 연구에서는 혼합모형의 케이스 제거 진단에 필요한 효율적인 계산방법을 제안하는 데 그 목적이 있다. Christensen, Pearson과 Johnson(1992), Hurtado(1993)의 논문은 개별 관측치를 제거한 후의 REMLE나 MLE를 구하는 경우에 전체자료를 이용하여 분산행렬의 역행렬을 먼저 구하고 계산된 역행렬의 행을 재배치한 뒤 적당한 선형변환을 통하여 개별 관측치를 제거한 후의 분산행렬의 역행렬을 구하고, 이것을 이용하여 분산성분을 추정하는 것으로 서술되어 있다. 하지만 일반적으로 많이 사용되는 분산성분추정량인 최우추정량(MLE), 제한최우추정량(REMLE) 및 최소노옴불편추정량(MINQUE)을 구하는 경우에는 분산행렬의 역행렬에 관한 계산은 주로 Hemmerle와 Hartley(1973)의 W-변환을 사용하며 이 계산방법은 구태여 차수가 전체자료의 개수인 분산행렬의 역행렬 계산을 필요로 하지 않는다. 한편 Lee와 Kim(1988)은 W-변환보다 더 효율적인 W-변환의 재구성 기법을 제안하였는데, 그 방법은 분산성분을 추정하는 경우에 분산행렬의 역행렬에 연관된 계산을 단지 비어있지 않는 셀의 개수를 차수로 하는 역행렬의 계산으로 바꿀 수 있다. 따라서 본 논문에서 제안되는 방법은 W-변환의 재구성 기법을 기초로 한 혼합모형의 케이스 제거진단에 필요한 효율적인 계산방법이며 이 방법의 기본착상은 반복이 있는 혼합모형을 반복이 없는 것처럼 활용할 수 있다는데 있다고 할 수 있다.

본 논문의 구성은 1절에서는 연구의 기본적인 배경을 설명하고, 2절에서는 혼합모형 및 분산성분추정량에 대하여 간략하게 소개한다. 연구의 핵심인 3절에서는 혼합모형의 케이스 제거진단에 필요한 REMLE, MLE 그리고 MINQUE와 같은 분산성분추정량을 구하는데 편리한 새로운 계산방법을 서술하며 4절에서는 3절의 계산방법을 비례적 분산의 경우로 확장한다. 그리고 5절에서는 예제를 통하여 SAS/IML을 이용하여 구하는 방법을 소개하며, 끝으로 6절에서는 본 연구의 결론이 주어진다.

2. 혼합모형 및 분산성분추정량

이 절에서는 혼합모형과 분산성분추정량의 종류에 대하여 언급하기로 한다. 혼합모형과 분산성분추정량에 대하여 보다 자세한 내용은 Searle(1971,1987)과 Rao와 Kleffe (1988)의 책에서 찾아볼 수 있다.

2.1 혼합모형

일반적으로 혼합모형은 다음과 같이 서술할 수 있는데,

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

여기서 $\underline{y}$는 알려진 $N\times 1$ 자료벡터, X는 계수 $r(X)=r(r<p)$인 고정효과에 관련된 $N\times p$ 계획행렬, Z는 랜덤효과에 관련된 $N\times q$ 계획행렬, β는 고정효과로 언급되는 $p\times 1$ 열벡터, γ는 랜덤효과로 언급되는 $q\times 1$ 열벡터, 그리고 $\underline{e}$는 $N\times 1$ 오차벡터이다. 또한 γ에 포함된 랜덤효과들에 대응되는 분산성분들을 표시하기 위하여 γ를 c개의 부분벡터 $\gamma' = (\underline{\gamma_1}' | \underline{\gamma_2}' | \cdots | \underline{\gamma_c}')$과 같이 분할하고 $\underline{\gamma_i}$에 대응되는 계획행렬을 Z_i로 두면 식(2.1)을 다음과 같은 식(2.2)로 표현할 수 있다.

$$\underline{y} = X\beta + Z_1\gamma_1 + Z_2\gamma_2 + \cdots + Z_c\gamma_c + \underline{e}. \tag{2.2}$$

식(2.2)에서 Z_i의 차수는 $N\times m_i$, γ_i는 $m_i\times 1$ 열벡터이며 $q = \sum_{i=1}^{c} m_i$이다. 아울러 식(2.2)의 분포에 관한 성질은 다음 식(2.3)과 같은 가정이 성립한다고 한다.

$$\gamma_i \sim N(0, \sigma_i^2 I_{m_i}),\quad \underline{e} \sim N(0, \sigma_e^2 I_N),\quad Cov(\gamma_i, \underline{e}') = 0,\quad Cov(\gamma_i, \gamma_j) = 0,\quad \forall\, i \neq j. \tag{2.3}$$

여기서 I_{m_i}은 차수가 $m_i\times m_i$인 항등행렬이며, I_N은 차수가 $N\times N$인 항등행렬 이고 따라서 $\underline{y}$는 평균이 $X\beta$이고 분산행렬이 $V = \sigma_e^2 I_N + \sum_{i=1}^{c} \sigma_i^2 Z_i Z_i'$인 다변량정규벡터이다.

위와 같은 분산행렬은 다음 식(2.4)와 같은 α-기호를 이용하여 주로 표기되는데,

$$\alpha_0 = \sigma_e^2, \quad \alpha_i = \sigma_i^2/\sigma_e^2. \tag{2.4}$$

행렬 H를 $V = \alpha_0 H$로 두면, H는 $H = I_N + \sum_{i=1}^{c} \alpha_i Z_i Z_i'$으로 나타내어 진다.

2.2 분산성분추정량

만일 분산행렬 V가 기지인 경우에는 $X\beta$의 최적추정량은 다음과 같이 구할 수 있다.

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

식(2.5)에서 분산행렬 V대신 행렬 H를 사용하여도 같음을 쉽게 알 수 있는데, 일반적으로 분산행렬 V 또는 행렬 H는 알 수가 없기 때문에, 주로 사용되는 방법은 주어진 자료를 이용하여 분

산성분 σ_i^2를 추정하고 추정된 분산성분의 값을 이용하여 식(2.5)에 대입하여 사용한다. 이 경우 오늘날 널리 사용되는 분산성분의 추정방법들은 헨더슨의 방법 III(1953), Hartley와 J.N.K.Rao (1967)의 최우추정량(MLE), C.R.Rao(1971)의 최소노옴불편추정량(MINQUE), 그리고 Patterson과 Thompson(1972)의 제한최우추정량(REMLE)등이 있다.

3. 혼합모형의 케이스 제거 진단에 대한 계산방법

혼합모형에 포함된 분산성분을 MLE, REMLE와 MINQUE등을 이용하여 추정하는 경우에는 $X'H^{-1}X$, $Z'H^{-1}Z$, $X'H^{-1}Z$, $X'H^{-1}\underline{y}$, $Z'H^{-1}\underline{y}$와 $\underline{y}'H^{-1}\underline{y}$와 같은 행렬연산을 하여야만 한다. 이와 같은 행렬연산을 한 개의 행렬로 표현한 식(3.1)과 같은 행렬을 Hemmerle와 Hartley(1973)는 W-행렬이라고 불렀다.

$$W = \begin{pmatrix} Z'H^{-1}Z & Z'H^{-1}X & Z'H^{-1}\underline{y} \\ X'H^{-1}Z & X'H^{-1}X & X'H^{-1}\underline{y} \\ \underline{y}'H^{-1}Z & \underline{y}'H^{-1}X & \underline{y}'H^{-1}\underline{y} \end{pmatrix}. \tag{3.1}$$

그리고 그들은 W-행렬을 쉽게 구할 수 있는 W-변환을 제안하였는데, 그 기본적인 아이디어는 $N\times N$ 행렬 H의 역행렬에 관한 계산을 단지 $q\times q$ 역행렬의 계산을 수행하여 구할 수 있다는 데 있다. 한편 Lee와 Kim(1988)은 W-행렬을 비어있지 않는 셀의 수를 n이라고 할 때 $n\times n$ 역행렬을 구하여 W-행렬을 구하는 W-행렬의 재구성 방법을 제안하였는데, 이 방법은 교호작용이 포함된 실험계획인 경우에는 대체적으로 n이 q보다 작으며, 큰 경우에는 간단한 행렬 변환식을 이용하여 항상 W-변환보다 계산방법이 효율적임을 알 수 있다. 다음은 Lee와 Kim(1988)의 결과를 요약, 정리한 것이다.

만일 반복 관측치가 존재하는 불균형계획행렬 X와 Z에 대한 각 셀 당 한 개의 관측치 만을 포함하는 행렬 X_0와 Z_0를 각각 비교하면 관계식이 $X = TX_0$와 $Z = TZ_0$가 되는 다음과 같은 반복행렬 T행렬이 존재함을 알 수 있다. 식(3.2)에서 n_i는 각 셀 당 포함된 관측치의 수, 1_{n_i}는 차수가 n_i인 1로 구성된 벡터, n은 혼합모형에서 비어있지 않는 셀의 개수를 의미하며, 반복행렬 T는 혼합모형에서 최고차 항의 교호작용에 해당되는 계획행렬이 된다.

$$T_{N\times n} = \begin{pmatrix} 1_{n_1} & 0 & \cdots & 0 \\ 0 & 1_{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1_{n_n} \end{pmatrix}. \tag{3.2}$$

한편 행렬 D를 $D = T'T$ 로 두면, D는 다음 식(3.3)과 같은 각 셀 당 반복의 개수를 표시하는 셀도수 행렬이 되는 사실을 쉽게 확인할 수 있다.

$$D_{n\times n} = \begin{pmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_n \end{pmatrix}. \tag{3.3}$$

그리고 관찰치벡터 $\underline{y}$에 대하여 셀평균벡터 $\bar{\underline{y}}$를 다음과 같이 정의하며, 이 경우 y_{ij}는 i번째 셀의

j번째 관측치를 의미한다.

$$\bar{\underline{y}} = \begin{pmatrix} \sum_{j=1}^{n_1} y_{1j}/n_1 \\ \sum_{j=1}^{n_2} y_{2j}/n_2 \\ \vdots \\ \sum_{j=1}^{n_n} y_{nj}/n_n \end{pmatrix}. \tag{3.4}$$

한편 행렬 U를 $U=[\sqrt{\alpha_1}Z_1|\sqrt{\alpha_2}Z_2|...|\sqrt{\alpha_c}Z_c]$로 두면, 행렬 H는 $H = I_N + UU'$로 표현되며, S^2을 $S^2 = \underline{y}'\underline{y} - \bar{\underline{y}}'D\bar{\underline{y}}$로 정의하면 식(3.1)의 W-행렬은 다음과 같은 행렬로 표현된다.

$$W^* = \begin{pmatrix} Z_0'MZ_0 & Z_0'MX_0 & Z_0'M\bar{\underline{y}} \\ X_0'MZ_0 & X_0'MX_0 & X_0'M\bar{\underline{y}} \\ \bar{\underline{y}}'MZ_0 & \bar{\underline{y}}'MX_0 & \bar{\underline{y}}'M\bar{\underline{y}}+S^2 \end{pmatrix}. \tag{3.5}$$

식(3.5)에 포함된 행렬 M은 $M = (I_n + DU_0U_0')^{-1}D$로 정의되며 여기서 I_n은 n×n인 항등행렬, U_0는 $U = TU_0$의 관계가 성립하는 균형계획행렬이다. 확실히 이 방법은 분산행렬의 역행렬 계산문제를 쉽게 취급할 수 있으며, 관측치벡터 $\underline{y}$에 관련된 계산을 셀평균벡터 $\bar{\underline{y}}$와 오차제곱합 S^2을 이용하여 계산하기 때문에 분산성분을 추정하는 데 매우 효율적인 계산방법을 제공한다.

이제 W-행렬의 재구성방법을 혼합모형에 대한 케이스 제거 진단에도 적용할 수 있음을 살펴보기로 한다. 이 경우 기본적인 생각은 일단 분산성분을 추정하기 위하여 M 행렬을 먼저 계산한 후에 이 행렬을 이용하여 각각의 개인 관측치를 제거한 다른 M 행렬을 어떻게 하면 쉽게 구할 수 있을까? 라는 질문으로 요약될 수 있다. 먼저 i번째 셀의 어떤 관측치를 제거하였다고 가정하여 보자. 그러면 이 경우의 셀도수행렬은 i번째 대각원소가 $n_i - 1$와 같이 바뀌게 되며 이 행렬을 D^*로 표기하자. 그리고 행렬을 간단하게 표기하기 위하여 행렬 P를 $P = (I_n + DU_0U_0')^{-1}$로 두면 행렬 M은 $M = PD$로 표현되며, i번째 셀의 어떤 관측치를 제거하고 구한 행렬 M을 M^*로 표기하면 $M^* = P^*D^*$, $P^* = (I_n + D^*U_0U_0')^{-1}$와 같다. 한편 행렬 P와 P^*에 대해서 간단한 행렬연산을 이용하면, $P^* = (I_n + P(P^{*-1} - P^{-1}))^{-1}P$가 성립함을 알 수 있다. 다음 보조정리는 행렬 P를 이용하여 P^*를 역행렬의 계산 없이 구하는데, 유용하게 사용된다.

[보조정리 3.1] (Graybill(1983), p.189)

k×k 행렬 $C = G + \delta\underline{a}\underline{b}'$에 대하여, 단 G는 정칙대각행렬, $\underline{a}$와 $\underline{b}$는 k×1 벡터, g_{ii}은 G의 i번째 대각원소, δ는 $\delta \neq -1/(\sum_{i=1}^{k} a_ib_i/g_{ii})$인 상수라고 하면, C의 역행렬은 다음과 같이 구할 수 있다.

$$C^{-1} = G^{-1} + \eta\underline{a}^*\underline{b}^{*\prime},\ \eta = -\delta(1+\delta\sum_{i=1}^{k} a_ib_ig_{ii}^{-1})^{-1},\ a_i^* = a_{ii}/g_{ii},\ b_i^* = b_{ii}/g_{ii}. \tag{3.6}$$

[정리 3.1] 행렬 M^*는 $M^* = (I + t\underline{v}_i\underline{u}_i')PD^*$으로 표현된다. 단 $\underline{u}_i'$는 행렬 U_0U_0'의 i번째 행, $\underline{v}_i$는 행렬 P의 i번째 열이며, 상수 t는 $t = 1/(1-\underline{u}_i'\underline{v}_i)$로 정의된다.

(증명) i번째 셀의 어떤 관측치를 제거하고 구한 행렬 $(P^{*-1}-P^{-1})$는 U_0U_0'의 i번째 행 $\underline{u}_i'$ 이외에는 모두 원소가 0인 $n\times n$ 행렬에 -를 붙인 행렬이다. 따라서 $P(P^{*-1}-P^{-1})$는 행렬 P의 i번째 열 $\underline{v}_i$와 행 $-\underline{u}_i'$의 곱으로 표현된다. 따라서 행렬 $I_n+P(P^{*-1}-P^{-1})$는 $I_n-\underline{v}_i\underline{u}_i'$와 같다. 그리고 이 행렬의 역행렬을 구하기 위하여 [보조정리 3.1]을 사용하면, 다음과 같음을 쉽게 확인할 수 있다.

$$(I_n+P(P^{*-1}-P^{-1}))^{-1}=(I_n+t\underline{v}_i\underline{u}_i'),\quad t=1/(1-\underline{u}_i'\underline{v}_i)$$

[정리 3.1]은 행렬 M^*를 구할 때 모든 관측치를 포함하고 구한 행렬 P를 어떻게 이용할 수 있는 지를 알려준다. 이제 남은 계산은 단지 i번째 셀의 어떤 관측치를 제거하고 구한 셀평균벡터와 오차제곱합 S^2만을 새롭게 계산하여 식(3.5)를 이용하면 특정 관측치를 제거하고 구한 W-행렬을 쉽게 구할 수 있다.

4. 비례분산인 경우에 대한 계산방법

이 절에서는 오차벡터 $\underline{e}$가 $\underline{e}'=[\underline{e}_1'|\underline{e}_2'|\cdots|\underline{e}_n']$와 같이 분할되는 경우에 3절에서 소개된 계산방법을 어떻게 사용할 수 있는 지를 설명하고자 한다. 여기서 부분벡터 $\underline{e}_1,\underline{e}_2,\ldots,\underline{e}_n$는 서로 독립이고, $i=1,2,\ldots,n$에 대해서 $\underline{e}_i$는 $\underline{e}_i\sim MVN(\underline{0},s_i^2I)$인 다변량 정규분포를 따른다고 가정한다. 만일 분산들이 l_i는 알려진 상수인 경우에 $s_i^2=l_i\sigma_e^2$의 관계식이 성립한다면, 모형(2.2)에 대한 분산행렬 V는 $V=\sigma_e^2H=\sigma_e^2(L+UU')$와 같이 표현되며, 이 경우 대각행렬 L은 i번째 대각블록행렬 L_{ii}가 $i=1,2,\ldots,n$에 대해서 $L_{ii}=l_iI_{n_i}$를 만족한다. 그러면 3절의 W-행렬의 재구성기법은 관측치벡터 $\underline{y}$에 대하여 $\underline{y}^\circ=L^{-1/2}\underline{y}$라고 두면 쉽게 적용이 가능하다.

이 경우 벡터 $\underline{y}^\circ$는 다변량정규분포를 따르며, 평균벡터와 분산행렬은 다음 식(4.1)과 같이 된다.

$$E(\underline{y}^\circ)=L^{-1/2}X\beta,\quad Var(\underline{y}^\circ)=\sigma_e^2(I+L^{-1/2}UU'L^{-1/2}).\tag{4.1}$$

그러면 W-행렬과 같은 W-행렬의 재구성기법을 사용하여 구한 W^{**}행렬은 관계식 $T'L^{-1/2}=L_0^{-1}T'$과 $\overline{\underline{y}}^\circ=L_0^{-1/2}\overline{\underline{y}}$을 이용하여 다음 식(4.2)과 같이 구할 수 있다. 이 경우 L_0행렬은 반복이 없는 경우의 L행렬이며, i번째 대각원소는 l_i로 표현된다.(Lee와 Kim(1988) 참조).

$$W^{**}=\begin{pmatrix} Z_0'QZ_0 & Z_0'QX_0 & Z_0'Q\overline{\underline{y}} \\ X_0'QZ_0 & X_0'QX_0 & X_0'Q\overline{\underline{y}} \\ \overline{\underline{y}}'QZ_0 & \overline{\underline{y}}'QX_0 & \overline{\underline{y}}'Q\overline{\underline{y}}+S^{*2}\end{pmatrix}.\tag{4.2}$$

식(4.2)에서 사용된 Q행렬은 $Q=(I_n+DL_0^{-1}U_0U_0')^{-1}DL_0^{-1}$로 정의되며, 상수 S^{*2}은 $S^{*2}=\underline{e}'L^{-1}\underline{e}$ 와 같이 정의된다.

위와 같은 비례분산인 경우에도 3절의 계산방법은 쉽게 적용할 수 있는데, i번째 셀의 어떤 관측치를 제거하였다고 가정하고, i번째 셀의 어떤 관측치를 제거하고 구한 행렬 Q를 Q^*로 표기하

면 $Q^* = R^*D^*L_0^{-1}$, $R^* = (I_n + D^*L_0^{-1}U_0U_0')^{-1}$와 같다. 따라서 3절의 계산방법을 따라서 Q^*를 구하는 방법은 다음 [정리 4.1]과 같이 약술된다.

[정리 4.1] 행렬 Q^*는 $Q^* = (I_n + t^*\underset{\sim}{y}_i\underset{\sim}{x}_i'/l_i)RD^*L_0^{-1}$으로 표현된다. 단 $\underset{\sim}{x}_i'$는 행렬 $L_0^{-1}U_0U_0'$의 i번째 행, $\underset{\sim}{y}_i$는 행렬 R의 i번째 열이며 상수 t^*는 $t^* = 1/(1-\underset{\sim}{x}_i'\underset{\sim}{y}_i)$, 그리고 행렬 R은 $R = (I_n + DL_0^{-1}U_0U_0')^{-1}$와 같이 정의된다.

(증명) [정리 3.1]의 풀이과정과 거의 흡사하다.

5. 예제

이 절에서는 3절에서 제안된 계산방법을 어떻게 활용할 수 있는 지를 실제자료를 이용하여 계산하는 방법을 알아본다. 다음 자료는 Bowker와 Lieberman(1963)의 실험에 의해 얻어진 자료이며 Hemmerle와 Hartley(1973)에 의하여 다시 사용된 것으로써 오븐의 종류와 온도에 따른 전자성분의 수명에 대한 자료이다.

온 도	오븐		
	1	2	3
1	237	208	186
	254	178	183
	246	187	
2	178	146	142
	179	145	125
		141	136

이 경우 오븐의 종류는 고정효과라고 생각하고 온도를 랜덤효과라고 생각하면 다음과 같은 이원혼합모형으로 생각할 수 있다.

$$y_{ijk} = \mu + a_i + b_j + c_{ij} + e_{ijk}$$

여기서 $i = 1, 2, 3$; $j = 1, 2$ 이고 b_j, c_{ij}, e_{ijk} 는 각각 평균이 0이고 분산이 σ_b^2, σ_c^2, σ_e^2인 정규분포를 따른다고 하자. 이 경우 일반성을 잃지 않고 예를 들어 온도는 2조건, 오븐이 1인 경우의 자료 179를 제거하고 W-행렬을 구하여 보기로 한다. 다음 <그림 5.1>은 W-행렬의 재구성 기법을 사용하는 경우에 필수적인 행렬 M^*, 새로 계산되는 셀평균벡터 $\bar{y}$와 오차제곱합 S^2를 구하는 과정을 보여주는 SAS/IML 프로그램이다.

위에서 고려된 자료의 전체 개수는 16개이므로, Christensen, Pearson과 Johnson(1992), Hurtado(1993)의 케이스 제거진단 계산방법을 이용하면, REMLE나 MLE를 구하기 위하여 16×16 역행렬을 계산하여야 하지만, 본 연구의 결과를 이용하면 단지 6×6 행렬의 역행렬 만을 고려하면 된다. 그리고 제안된 알고리즘은 위의 예에서 반복 수를 늘려서 자료가 10000개가 되는 경우에

도 10000×10000 역행렬을 구하지 않고, 단지 6×6 행렬의 역행렬 만을 고려하면 되기 때문에 자료의 개수가 많으면 많을수록 효율성이 더욱 증가함을 알 수 있다.

<그림 5.1> 데이터에 대한 SAS/IML 프로그램

```
PROC IML;
RESET NOLOG;

/* 관찰치 입력 */
Y11={237, 254, 246}; Y12={178, 179}; Y21={208, 178, 187};
Y22={146, 145, 141}; Y31={186, 183}; Y32={142, 125, 136};

/* 관찰치 벡터 Y */
Y=Y11//Y12//Y21//Y21//Y31//Y32;

/* 총자료의 개수 */
TN=NROW(Y);

/* 셀평균 벡터 */
YBAR=Y11[:,]//Y12[:,]//Y21[:,]//Y22[:,]//Y31[:,]//Y32[:,];

/* 셀도수 행렬 */
DD=NROW(Y11)//NROW(Y12)//NROW(Y21)//NROW(Y22)//NROW(Y31)//NROW(Y32);
D=DIAG(DD);

/* 상수 S2의 계산 */
SSE=Y`*Y-YBAR`*D*YBAR;

/* 관측치 179를 제거한 후의 관찰치 벡터 */
NY12={178};
NY=Y11//NY12//Y21//Y21//Y31//Y32;

/* 관측치 179를 제거한 후의 셀평균 벡터 */
NYBAR=Y11[:,]//NY12[:,]//Y21[:,]//Y22[:,]//Y31[:,]//Y32[:,];

/* 관측치 179를 제거한 후의 셀도수 행렬 */
NDD=NROW(Y11)//(NROW(Y12)-1)//NROW(Y21)//NROW(Y22)//NROW(Y31)//NROW(Y32);
ND=DIAG(NDD);

/* 관측치 179를 제거한 후의 상수 S2의 계산 */
NSSE=NY`*NY-NYBAR`*ND*NYBAR;

/* 비어 있지 않은 셀의 개수 */
N=NROW(D);

/* 랜덤효과들에 대한 계획행렬 */
A1={1,2,1,2,1,2};
ZO1=DESIGN(A1); ZO2=I(N);

/* MINQUE(1), REMLE, MLE를 위한 초기값 */
K1=1; K2=1;

/* 전체 자료를 이용하여 구한 M 행렬 */
UO1=ZO1#SQRT(K1); UO2=ZO2#SQRT(K2); UO=UO1||UO2; UP=UO*UO`;
P=INV(I(N)+D*UO*UO`); M=P*D;

/* 관측치 179를 제거한 후의 M 행렬: 본 논문의 알고리즘을 이용하여 작성 */
U2=UP[2,]; V2=P[,2];
CONST=1/(1-U2*V2);
NM=(I(N)+CONST#V2*U2)*P*ND;

QUIT;
```

6. 결론

본 연구에서는 혼합모형의 케이스 제거진단에 필요한 분산성분추정량을 구하는데 사용될 수 있는 편리한 계산방법을 제안하였으며, 고려된 계산방법은 기존의 선행연구보다 훨씬 효율적임을 알 수 있다. 하지만 혼합모형에 있어서 고정효과의 추정, 변량효과의 예측, 분산성분의 추정, 그리고 혼합모형에 있어서 관심의 대상이 될 수 있는 고전적인 다른 모수의 추정에 관하여 관측치의 영향력을 계산하는데 사용되는 통계적 측도를 구하는 문제는 취급하지 못하였다. 지금까지 진행된 연구결과들은 거의 대부분 회귀진단에서 사용되는 통계적 측도들을 사용하는 것이 보편화되어 있으나, 역시 한정된 몇 가지 진단 측도 들에 국한되어 있다. 따라서 보다 효율적인 혼합모형의 진단을 위하여 회귀모형에서 사용되는 여러 가지 영향력관찰점의 측정도구 중 가장 합리적인 진단도구들을 찾는 것도 앞으로 연구되어져야 할 중요한 연구과제라고 할 수 있다.

참 고 문 헌

[1] Belsley, D. A., Kuh, E., Welsch, R. E. (1982). *Regression Diagnostics,* New York: John Wiley.

[2] Bowker, A. H. and Lieberman, G. H. (1963). *Engineering Statistics,* Englewood Cliffs, N.J., Prentice-Hall.

[3] Christensen, R. , Pearson, L. M. and Johnson, W. (1992). Case-Deletion Diagnostics for Mixed Models, *Technometrics,* Vol.34(1), 38-44.

[4] Cook, R. D. (1977). Detection of Influential Observations in Linear Regression, *Technometric* , 19, 15-18.

[5] Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression,* New York: Chapman and Hall.

[6] Fellner, W. H. (1986). Robust Estimation of Variance Components, *Technometrics,* 28, 51-60.

[7] Graybill, F. A. (1983). *Matrices with Applications in Statistics,* Wadsworth, Inc.

[8] Hartley, H. O. and Rao, J. N. K. (1967). Maximum Likelihood Estimation for the Mixed Analysis of Variance Model, *Biometrika,* 54, 93-108.

[9] Hemmerle, W. J. and Hartley, H. O. (1973). Computing Maximum Likelihood Estimates for the Mixed A.O.V. model Using the W Transformation, *Technometric,* 18, 207-211

[10] Henderson, C. R. (1953). Estimation of Variance and Covariance Components, *Biometrics,* 9, 226-252.

[11] Hocking, R. R. (1983). A Diagnostic Tool for Mixed Models with Application to Negative Estimates of Variance Components, *SUGI 8,* 711-716.

[12] Hocking, R. R. and Bremer, R. H. (1987). Estimation of Variance Components in Mixed

Factorial Models Including Model-Based Diagnostics, *SUGI 12,* 1162-1167.

[13] Hurtado, G. (1993). *Detection of Influential Observations in Linear Mixed Models,* Ph.D. dissertation, North Carolina State University.

[14] Lee, J. T. and Kim, B. C. (1988). A New Approach for the W-matrix, *Journal of Statistical Computation and Simulation,* 29, 241-254.

[15] Patterson, H. D. and Thompson, R. (1971). Recovery of Interblock Information When Block Sizes are Unequal, *Biometrika,* 58, 545-554.

[16] Rao, C. R. (1971). Estimation of Variance Components - MINQUE Theory, *Journal of Multivariate Analysis,* Vol.1, 257-275.

[17] Rao, C. R. and Kleffe, J. (1988). *Estimation of Variance Components and Applications,* Amsterdam: North-Holland.

[18] Searle, S. R. (1971). *Linear Models,* John Wiley & Sons, New York.

[19] Searle, S. R. (1987). *Linear Models for Unbalanced Data,* John Wiley & Sons, New York.

[20] 최병선 (1997). 「회귀분석(上)」 세경사.