Journal of the Korean Data & Information Science Society
2011, **22**(3), 381–387

한국데이터정보과학회지

사영을 이용한 일원 분산성분

최재성[1]

[1]계명대학교 통계학과

접수 2011년 3월 2일, 수정 2011년 4월 22일, 게재확정 2011년 4월 27일

요 약

본 논문은 일원 확률모형의 가정하에 실험자료를 분석할 때 확률모형과 관련된 분산성분을 추정하는 문제를 다루고 있다. 분산성분의 추정방법으로 적률법을 이용하고 있다. 적률법을 이용할 때 필요한 두 가지 계산과정은 요인의 변동에 따른 제곱합과 제곱합의 기대값 계산이다. 제곱합의 계산으로 사영을 어떻게 이용하는 가를 논의하고 있다. 제곱합의 기대값 계산을 위해 분산성분의 계수로 관측되는 관련행렬의 고유근을 이용하는 방법을 다루고 있다. 분산성분의 적률추정량으로 사영과 고유근을 이용한 분산성분의 추정방법이 Hartley (1967)의 합성법보다 간편하고 효율적인 방법임을 논의하고 있다.

주요용어: 고유근, 분산성분, 사영, 적률법, 확률모형.

1. 서론

실험자료를 분석하기 위한 모형으로 선형모형을 가정할 때, 실험의 성격 또는 특성에 따라 다양한 분석모형의 가정하에 자료를 분석하게 된다. 자료분석에 이용되는 모형은 개체 또는 실험단위의 반응에 영향을 미치는 요인들과 실험설계에 따라 달라진다.

Milliken과 Johnson (1984)은 실험계획을 처리구조와 설계구조의 두 구조로 다루고 있다. 처리구조는 실험자가 연구 또는 비교를 위해 선정한 처리, 처리조합 또는 모집단의 모임으로 정의하고 있다. 설계구조는 실험단위들의 동질적인 그룹 또는 블록들로 구성됨을 의미하고 있다. 비교하고자 하는 처리들이 요인들의 수준결합으로 구성될 때, 요인의 수가 둘이면 두 요인의 수준결합으로 주어지는 처리조합이 실험단위들에 배정되는 설계구조하에 실험이 행해짐을 의미한다. 요인의 수가 셋이면 세 요인의 수준결합으로 주어지는 처리조합이 처리로 간주됨을 의미한다. 비교하고자 하는 처리들이 요인들의 수준결합이 아닌 경우에 처리구조는 일원구조이다.

처리구조를 구성하는 요인으로 고정요인과 확률요인으로 구분할 수 있다. 처리구조를 나타내는 요인들이 모두 확률요인일 때, 가정된 선형모형은 확률모형 또는 확률효과모형으로 간주된다. 자료분석에 가정된 모형이 확률모형일 때 분석의 목적은 분산성분들에 대한 추론이다. 확률모형의 가정하에 분산성분들의 추론방법에 대한 많은 연구가 행해져 왔다. Searle (1971)과 Graybill (1976)은 여러 확률모형에 대한 분산성분의 적률추정량과 관련된 성질들을 다루고 있다. 이외에도 처리구조에 반복측정요인이 포함될 때 Galecki (1994)는 공분산구조를 다루고 있고 Choi (2008a, 2008b, 2010)은 모형과 분석방법을 제시하고 있다. McCullagh와 Nelder (1989)는 일반화된 선형모형에서 확률성분의 추론방법을 논의하고 있다.

[1] (704-701) 대구광역시 달서구 신당동 1000번지, 계명대학교 통계학과, 교수. E-mail: jschoi@kmu.ac.kr

본 논문은 확률모형내 관심모수인 분산성분들의 여러 추정방법중 적률법 (method of moments)에 의한 추정방법을 사용하기로 한다. 적률법은 관측변수의 분포에 대한 가정없이 불편성의 성질을 갖는 분산성분의 적률추정량을 제공하며 다른 방법들에 비해 상대적으로 계산하기에 편리한 이점이 있다. 적률법으로 분산성분을 추정하기 위해 모형내 내포된 모수들과 관련된 제곱합의 계산과 제곱합의 기대값에서 주어지는 분산성분들의 계수를 계산해야 한다.

분산성분들의 계수계산에 이용될 수 있는 방법은 두 가지이다. 하나는 대수적 방법 (algebraic method)이고 다른 하나는 Hartley (1967)에 의한 합성법 (synthesis)이다. 대수적 방법은 확률모형의 가정을 이용하여 대수적으로 제곱합의 기대값을 계산한다. 합성법은 제곱합의 기대치 계산으로 컴퓨터 기법을 적용하고 있다. 본 논문에서는 합성법을 이용하되 기존의 복잡한 계산방식을 단순하게 하는 사영에 의한 분산성분의 추정방법을 제공하고자 한다. 사영에 대한 구체적 논의는 Johnson과 Wichern (1988) 그리고 Graybill (1983) 등에서 살펴볼 수 있다.

2. 모형에 대한 가정

실험자료를 분석하기 위한 모형으로 일원 확률모형을 가정한다. 실험단위의 반응에 영향을 미치는 처리요인 A의 a개 수준은 수준들의 모집단에서 임의로 추출된다고 하자. 따라서 확률요인 A의 i번째 수준효과 α_i $(i = 1, 2, \cdots, a)$는 확률효과를 나타낸다. 요인 A의 a개 수준들은 n_i개 실험단위들에 임의로 배정된다고 가정한다. 요인 A의 수준 i에서 j번째 실험단위의 관측값을 y_{ij}라 둘 때 확률모형은 다음과 같다.

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{2.1}$$

단, μ는 전체평균이고 α_i는 평균이 0이고 분산이 σ_A^2이다. 오차항 ϵ_{ij}는 평균이 0이고 분산이 σ_ϵ^2이라 가정한다. 일원 구조의 확률모형에 내포된 모수들은 평균 μ와 두 개의 분산성분 σ_A^2과 오차성분 σ_ϵ^2이다. 관측값 y_{ij}에 대한 확률효과모형을 행렬로 표현하면 다음과 같다.

$$\boldsymbol{y} = \boldsymbol{j}\mu + \boldsymbol{X}_A\boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{2.2}$$

단, $\boldsymbol{y}$는 $n = \sum_{i=1}^{a} n_i$일 때 $n \times 1$인 관측벡터이다. $\boldsymbol{j}$는 모든 원소가 1인 $n \times 1$벡터이다. $\boldsymbol{X}_A$는 $n \times a$인 확률효과의 계수행렬을 나타낸다. $\boldsymbol{\epsilon}$은 $n \times 1$인 오차벡터이고 다변량 정규분포 $N(\boldsymbol{0}, \sigma_\epsilon^2\boldsymbol{I_n})$이며, 확률효과벡터 $\boldsymbol{\alpha}$는 $a \times 1$인 벡터로 $N(\boldsymbol{0}, \sigma_A^2\boldsymbol{I_a})$를 따른다고 가정한다. 관측벡터 $\boldsymbol{y}$의 공분산행렬을 $\boldsymbol{\Sigma}$로 둘 때,

$$\begin{aligned}\boldsymbol{\Sigma} &= var(\boldsymbol{y}) = var(\boldsymbol{j}\mu + \boldsymbol{X}_A\boldsymbol{\alpha} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{X}_A var(\boldsymbol{\alpha})\boldsymbol{X}_A' + var(\boldsymbol{\epsilon}) \\ &= \sigma_A^2\boldsymbol{X}_A\boldsymbol{X}_A' + \sigma_\epsilon^2\boldsymbol{I_n}\end{aligned} \tag{2.3}$$

이다. 따라서 단일요인 A의 확률효과모형에서 추정해야 할 두 분산성분은 σ_A^2과 σ_ϵ^2임을 알 수 있다.

3. 사영에 의한 분산성분

행렬모형식 (2.2)로부터 두 분산성분을 추정하기 위한 방법을 생각해 보기로 한다. 분산성분을 구하는 방법으로 적률법을 포함한 최대우도법과 MINQUE 방법을 생각할 수 있으나 분포에 대한 가정없이 분산성분을 추정할 수 있는 적률법을 이용하기로 한다. 적률법의 이점은 다른 방법들에 비해 계산이 간편하다는 점이다. 적률법에 의한 추정량은 불편성의 성질을 갖는다. 분산성분을 구하기 위한 첫 단계는

모형내 포함되어 있는 분산성분들의 수만큼 제곱합을 계산하는 것이다. 제곱합의 계산방법으로 일반적으로 일원분류하의 분산분석법을 적용한다. 분산분석법의 적용에서 두 분산성분 σ_A^2과 σ_ϵ^2에 해당하는 제곱합을 Q_A와 Q_ϵ라 두자.

$$Q_A = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{y_{i.}} - \bar{y_{..}})^2 \text{ 이고 } Q_\epsilon = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y_{i.}})^2 \text{이다.}$$

단, $y_{i.} = \sum_{j=1}^{n_i} y_{ij}$이고 $y_{..} = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}$이다. $\bar{y_{i.}}$는 확률요인 A의 수준 i에서 관측값들의 평균이고 $\bar{y_{..}}$는 자료들의 전체평균을 나타낸다. Q_A와 Q_ϵ을 분산분석에서 적용되는 대수적 방법이 아닌 사영행렬로 구해보기로 한다. 일차원상의 관측자료를 n차원상의 공간벡터로 나타낼 때, 다변량 벡터에 대한 행렬모형식 (2.2)를 이용할 수 있다. 식 (2.2)에서 모형행렬을 $\boldsymbol{X}$라 둘 때, $\boldsymbol{X} = (\boldsymbol{j},\ \boldsymbol{X}_A)$이다. 즉, 식 (2.2)는

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{\epsilon} \\ &= (\boldsymbol{j}, \boldsymbol{X}_A)\begin{pmatrix}\mu \\ \boldsymbol{\alpha}\end{pmatrix} + \boldsymbol{\epsilon} \\ &= \boldsymbol{j}\mu + \boldsymbol{X}_A\boldsymbol{\alpha} + \boldsymbol{\epsilon} \qquad (3.1)\end{aligned}$$

임을 나타낸다. 모형행렬 $\boldsymbol{X}$에 의해 생성된 벡터공간을 χ라 둘 때, 벡터공간 χ는 고정효과의 추정을 위한 계수벡터 $\boldsymbol{j}$에 의해 생성된 벡터공간과 확률효과들의 계획행렬로 생성되는 벡터공간의 두 부분공간으로 구성된다. 모형행렬로 주어지는 벡터공간 χ내 두 부분벡터공간은 직교하는 벡터공간이다. 그 이유는 식 (3.1)은 $\boldsymbol{y} = \boldsymbol{j}\mu + \boldsymbol{\epsilon}'$로 표현될 수 있고 $\boldsymbol{\epsilon}' = \boldsymbol{X}_A\boldsymbol{\alpha} + \boldsymbol{\epsilon}$를 나타내고 있기 때문이다.

사영에 의한 제곱합을 생각해 보자. 관측벡터 $\boldsymbol{y}$를 계수벡터 $\boldsymbol{j}$로 생성된 벡터공간으로의 사영은 $\boldsymbol{jj}^-\boldsymbol{y}$로 구해진다. 관측벡터 $\boldsymbol{y}$를 $\boldsymbol{j}$로의 사영에 의한 제곱거리를 SSM이라 두자. $SSM = \boldsymbol{y}'\boldsymbol{jj}^-\boldsymbol{y}$이다. 단, $\boldsymbol{j}^- = (\boldsymbol{j}'\boldsymbol{j})^{-1}\boldsymbol{j}'$이다. 확률요인 A에 따른 제곱합을 SSA라 두자. SSA를 구하기 위해 관측벡터 $\boldsymbol{y}$를 $\boldsymbol{X}_A$로의 사영을 생각해 보자. 확률효과 벡터 $\boldsymbol{\alpha}$의 계획행렬 $\boldsymbol{X}_A$에 의해 생성된 벡터공간으로의 사영을 나타내는 사영행렬은 $\boldsymbol{X}_A\boldsymbol{X}_A^-$이다. 모형행렬 $\boldsymbol{X} = (\boldsymbol{j},\ \boldsymbol{X}_A)$에 의해 생성된 벡터공간 χ는 상호직교하는 두 부분벡터 공간으로 구성되므로 $\boldsymbol{X}_A$에 의해 생성된 벡터공간으로의 사영행렬은 $\boldsymbol{X}_A\boldsymbol{X}_A^- = \boldsymbol{XX}^- - \boldsymbol{jj}^-$로 구해진다. SSA를 구하기 위해 관측벡터 $\boldsymbol{y}$를 $\boldsymbol{j}$로의 사영 $\boldsymbol{jj}^-\boldsymbol{y}$를 제외한 성분벡터를 $\boldsymbol{y_1}$이라 두자. 즉,

$$\begin{aligned}\boldsymbol{y_1} &= \boldsymbol{y} - \boldsymbol{jj}^-\boldsymbol{y} \\ &= (\boldsymbol{I} - \boldsymbol{jj}^-)\boldsymbol{y}\end{aligned}$$

로 표현된다. 따라서, 관측벡터 $\boldsymbol{y}$를 $\boldsymbol{j}$로의 사영을 제외한 편차벡터 $\boldsymbol{y_1}$을 $\boldsymbol{X}_A$로의 사영은 $\boldsymbol{X}_A\boldsymbol{X}_A^-\boldsymbol{y}$ 또는 $(\boldsymbol{XX}^- - \boldsymbol{jj}^-)\boldsymbol{y}$로 구해진다. SSA는 $\boldsymbol{y_1}$을 $\boldsymbol{X}_A$로의 사영에 이르는 제곱거리를 나타내므로 $SSA = \boldsymbol{y_1}'(\boldsymbol{XX}^- - \boldsymbol{jj}^-)\boldsymbol{y_1}$이 된다. SSA는 또한 관측벡터 $\boldsymbol{y}$의 2차형식으로도 계산됨을 알 수 있다. 즉, $SSA = \boldsymbol{y}'(\boldsymbol{XX}^- - \boldsymbol{jj}^-)\boldsymbol{y}$로 계산된다. 오차제곱합을 SSE라 두자. SSE는 관측벡터 $\boldsymbol{y}$를 오차벡터 $\boldsymbol{\epsilon}$로의 사영이 행해질 때, 사영까지의 제곱거리를 나타낸다. 사영행렬은 $\boldsymbol{I} - \boldsymbol{XX}^-$이고 제곱거리는 $\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{XX}^-)\boldsymbol{y}$로 주어지므로 $SSE = \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{XX}^-)\boldsymbol{y}$이다. 단일요인의 확률모형의 가정하에 행해진 관측벡터 $\boldsymbol{y}$의 총제곱합을 TSS라 둘때 $TSS = \boldsymbol{y}'\boldsymbol{y}$이다. 총제곱합 TSS는 다음과 같이 분해된다.

$$\boldsymbol{y}'\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{jj}^-\boldsymbol{y} + \boldsymbol{y}'(\boldsymbol{XX}^- - \boldsymbol{jj}^-)\boldsymbol{y} + \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{XX}^-)\boldsymbol{y} \qquad (3.2)$$

식 (3.2)는 $TSS = SSM + SSA + SSE$임을 나타내고 있다. 적률법에 의해 분산성분 추정량을 구할 때의 핵심은 변동요인에 따른 제곱합의 계산과 평균평방합의 기대값을 파악하는 것이 된다. 식 (3.2)에서 각 성분들의 제곱합은 관측벡터 $\boldsymbol{y}$의 2차형식으로 주어지므로 2차형식의 기대값을 생각해 보자. 행렬 $\boldsymbol{A}$가 2차형식의 행렬일 때 기대값은

$$E(\boldsymbol{Y}'\boldsymbol{AY}) = tr(\boldsymbol{A\Sigma}) + \mu^2\boldsymbol{j}'\boldsymbol{Aj} \tag{3.3}$$

으로 구해진다. 식 (3.2)에서 주어지는 2차형식의 행렬은 상호 직교하는 벡터공간으로의 사영을 나타내는 사영행렬이므로 $\mu^2\boldsymbol{j}'\boldsymbol{Aj} = 0$임을 알 수 있다. 따라서, 식 (3.3)은 $E(\boldsymbol{Y}'\boldsymbol{AY}) = tr(\boldsymbol{A\Sigma})$로 구해진다. 확률모형의 가정하에 주어진 $\boldsymbol{\Sigma}$는 식 (2.3)에서 $\boldsymbol{\Sigma} = \sigma_A^2\boldsymbol{X}_A\boldsymbol{X}_A' + \sigma_\epsilon^2\boldsymbol{I_n}$이다. 그러므로 식 (3.3)에 적용하면

$$E(\boldsymbol{Y}'\boldsymbol{AY}) = \sigma_A^2 tr(\boldsymbol{AX}_A\boldsymbol{X}_A') + \sigma_\epsilon^2 tr(\boldsymbol{A}) \tag{3.4}$$

로 구해진다. 2차형식의 기대값을 구하는 방법으로는 식 (2.1)의 가정을 이용한 대수적 방법이나 Hartley (1967)의 합성법 (synthesis)이 이용된다. Hartley (1967)는 확률성분의 계수인 대각합 (trace)을 구하기 위한 방법을 제공하고 있으나 계산이 간편하지 않다. 2차형식의 행렬들은 사영행렬로 주어지고 사영행렬은 멱등행렬이므로 멱등행렬의 성질을 이용할 수 있다.

본 논문에서는 식 (3.4)에 나타나는 행렬의 대각합이 고유근의 합이 된다는 점을 이용하여 행렬 $\boldsymbol{AX}_A\boldsymbol{X}_A'$의 고유근의 합과 행렬 A의 고유근의 합을 이용한다. 고유근을 이용할 때, 식 (3.4)는 다음과 같다.

$$E(\boldsymbol{Y}'\boldsymbol{AY}) = \sigma_A^2\sum_{i=1}^{n}\lambda_i(\boldsymbol{AX}_A\boldsymbol{X}_A') + \sigma_\epsilon^2\sum_{i=1}^{n}\lambda_i(\boldsymbol{A}) \tag{3.5}$$

단, $\lambda_i(\cdot)$는 $(\cdot)$내 행렬의 고유근을 나타낸다. 식 (3.5)에 의한 평균평방합의 기대값을 구해보기로 한다. 확률요인 A의 변동량을 나타내는 Q_A의 기대값과 오차에 따른 변동량을 나타내는 Q_ϵ의 기대값을 식 (3.5)로 구하면 다음과 같이 주어진다.

$$E(Q_A) = \sigma_A^2\sum_{i=1}^{n}\lambda_{iA(Q_A)}((\boldsymbol{XX}^- - \boldsymbol{jj}^-)\boldsymbol{X}_A\boldsymbol{X}_A') + \sigma_\epsilon^2\sum_{i=1}^{n}\lambda_{i\epsilon(Q_A)}(\boldsymbol{XX}^- - \boldsymbol{jj}^-) \tag{3.6}$$
$$E(Q_\epsilon) = \sigma_A^2\sum_{i=1}^{n}\lambda_{iA(Q_\epsilon)}((\boldsymbol{I} - \boldsymbol{XX}^-)\boldsymbol{X}_A\boldsymbol{X}_A') + \sigma_\epsilon^2\sum_{i=1}^{n}\lambda_{i\epsilon(Q_\epsilon)}(\boldsymbol{I} - \boldsymbol{XX}^-)$$

단, $\lambda_{iA(Q_A)}$는 Q_A의 기대값 계산에서 분산성분 σ_A^2의 계수를 구하기 위한 관련행렬의 i번째 고유근을 나타낸다. $\lambda_{i\epsilon(Q_A)}$은 Q_A의 기대값 계산에서 분산성분 σ_ϵ^2의 계수를 구하기 위한 관련행렬의 i번째 고유근을 나타낸다. $\lambda_{iA(Q_\epsilon)}$는 Q_ϵ의 기대값 계산에서 분산성분 σ_A^2의 계수를 구하기 위한 관련행렬의 i번째 고유근을 나타낸다. $\lambda_{i\epsilon(Q_\epsilon)}$은 Q_ϵ의 기대값 계산에서 분산성분 σ_ϵ^2의 계수를 구하기 위한 관련행렬의 i번째 고유근을 나타낸다. 로 주어지는 식 (3.6)의 $E(Q_\epsilon)$의 우측 첫 항은 0이 된다. 적률법에 의한 두 분산성분의 추정량들을 $\widehat{\sigma_A^2}$, $\widehat{\sigma_\epsilon^2}$이라 두자. $\widehat{\sigma_A^2}$과 $\widehat{\sigma_\epsilon^2}$은 다음 두 연립방정식의 해로 구해진다.

$$\boldsymbol{y}'(\boldsymbol{XX}^- - \boldsymbol{jj}^-)\boldsymbol{y} = \widehat{\sigma_A^2}\lambda_{A(Q_A)} + \widehat{\sigma_\epsilon^2}\lambda_{\epsilon(Q_A)} \tag{3.7}$$
$$\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{XX}^-)\boldsymbol{y} = \widehat{\sigma_\epsilon^2}\lambda_{\epsilon(Q_\epsilon)}$$

단, $\lambda_{A(Q_A)}$는 양반정치행렬 $(\boldsymbol{XX}^- - \boldsymbol{jj}^-)\boldsymbol{X}_A\boldsymbol{X}_A'$의 고유근들의 합이다. $\lambda_{\epsilon(Q_A)}$는 $(\boldsymbol{XX}^- - \boldsymbol{jj}^-)$의 고유근들의 합을 나타낸다. $\lambda_{\epsilon(Q_\epsilon)}$은 멱등행렬 $\boldsymbol{I} - \boldsymbol{XX}^-$의 고유근들의 합을 나타낸다. 식 (3.7)은 확률요인이 하나인 확률모형의 가정하에 분산성분을 추정하기 위한 적률방법을 이용할 때, 사영행렬과 고유근을 이용하는 것이 Hartley (1967)의 합성법보다 계산이 단순하고 효율적임을 나타내고 있다.

4. 분산성분의 계산 예

표 4.1은 밀의 네가지 품종에 대한 자료 예이다. 자료는 Milliken과 Johnson (1984)이 적률법으로 일원 확률모형의 가정하에 분산성분을 추정하는 데 사용한 자료를 나타낸다. 동일 자료에 대한 분산성분의 추정값이 식 (3.7)의 사영행렬과 고유근에 의해서도 구해진다. 밀품종의 모집단에서 임의로 추출된 네 품종이 품종별로 최대 네 개의 실험구를 갖는 완전임의 배열법에 의해 실험이 행해지고 실험구당 20개 작물이 임의로 추출되어 추출작물당 해충에 의한 피해량이 0에서 10까지의 척도로 관측된 자료이다.

표 4.1 네 가지 밀 품종에 대한 해충 피해량의 자료

품종A	품종B	품종C	품종D
3.90	3.60	4.15	3.35
4.05	4.20	4.60	3.80
4.25	4.05	4.15	
	3.85	4.40	

품종A에 임의로 세 개의 실험구가 배정되고 이중 한 실험구에서 임의로 선택된 20개 작물에서 측정된 해충의 피해량이 3.90임을 나타내고 있다. 밀 품종의 모집단에서 임의로 네 품종이 추출되었기 때문에 추출된 네 품종은 확률표본으로 간주되고 이들 처리효과는 확률효과로 $N(0, \sigma_A^2)$인 분포를 따른다고 가정한다. 따라서, 식 (2.1)의 확률모형의 가정하에 분산추정치를 구해본다. 식 (3.2)에 해당하는 계산량을 살펴보자. TSS 또는 $\boldsymbol{y}'\boldsymbol{y} = 212.1275$이고 SSM 또는 $\boldsymbol{y}'\boldsymbol{j}\boldsymbol{j}^-\boldsymbol{y} = 210.8094$로 구해진다. Q_A 또는 SSA로 표현된 $\boldsymbol{y}'(\boldsymbol{X}\boldsymbol{X}^- - \boldsymbol{j}\boldsymbol{j}^-)\boldsymbol{y} = 0.8101603$이다. Q_ϵ 또는 SSE로 표현된 $\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^-)\boldsymbol{y} = 0.5079167$로 주어진다. 총제곱합 TSS는 부분 벡터공간으로의 사영까지의 제곱거리의 합으로 구해짐을 알 수 있다.

식 (3.6)에서 제곱합 Q_A 또는 SSA의 기대값을 구하기 위해 관련된 두 행렬 $(\boldsymbol{X}\boldsymbol{X}^- - \boldsymbol{j}\boldsymbol{j}^-)\boldsymbol{X}\boldsymbol{X}'$와 $\boldsymbol{X}\boldsymbol{X}^- - \boldsymbol{j}\boldsymbol{j}^-$의 고유근들을 구해야 한다. $(\boldsymbol{X}\boldsymbol{X}^- - \boldsymbol{j}\boldsymbol{j}^-)\boldsymbol{X}\boldsymbol{X}'$는 13×13인 정방행렬이므로 13개의 고유근이 존재한다. 이 중 0아닌 세 고유근들은 4.000000, 3.302169, 2.236292이고 나머지 10개의 고유근들은 0의 값을 취한다. 0아닌 고유근들의 합 $\lambda_{A(Q_A)}$=9.538462로 구해진다. 다음에 $\boldsymbol{X}\boldsymbol{X}^- - \boldsymbol{j}\boldsymbol{j}^-$의 고유근들을 구해보자. $\boldsymbol{X}\boldsymbol{X}^- - \boldsymbol{j}\boldsymbol{j}^-$는 멱등행렬이므로 고유근은 0 아니면 1의 값을 취한다. $\boldsymbol{X}\boldsymbol{X}^- - \boldsymbol{j}\boldsymbol{j}^-$의 고유근중 세 개는 1이고 나머지는 0의 값으로 주어진다. 따라서 $\lambda_{\epsilon(Q_A)}$=3이다.

식 (3.6)의 두 번째 방정식에서 주어지는 $\lambda_{iA(Q_\epsilon)}$는 관련행렬 $(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^-)\boldsymbol{X}_A\boldsymbol{X}_A'$의 고유근을 나타낸다. $(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^-)\boldsymbol{X}_A\boldsymbol{X}_A'$의 고유근은 모두 0이므로 $\lambda_{\epsilon(Q_A)}$=0이다. $\lambda_{i\epsilon(Q_\epsilon)}$은 $\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^-$의 고유근이므로 0과 1로 관측된다. 13개의 고유근중 9개의 고유근은 1이고 4개는 0으로 관측된다. 따라서 고유근의 합 $\lambda_{\epsilon(Q_\epsilon)}$=9이다. 식 (3.7)의 연립방정식은 다음과 같이 주어진다.

$$\begin{aligned} 0.810603 &= 9.538462\widehat{\sigma_A^2} + 3\widehat{\sigma_\epsilon^2} \\ 0.5079167 &= 9\widehat{\sigma_\epsilon^2} \end{aligned} \tag{4.1}$$

식 (4.1)의 해로 $\widehat{\sigma_\epsilon^2} = 0.05643519$이고 $\widehat{\sigma_A^2} = 0.06718638$로 계산된다.

5. 결론

본 논문은 실험자료의 분석모형으로 일원 확률모형을 가정하고 있다. 실험에 이용되는 실험단위들은 동질적이다는 가정하에 실험 설계구조로 완전임의배열법을 가정한다. 단일 요인의 확률효과를 포함하는 확률모형내의 분산성분을 추정하기 위한 방법으로 적률법을 이용하고 있다. 적률법으로 분산추정을

할 때, 중요한 계산과정은 확률요인의 변동에 따른 제곱합의 계산과 제곱합의 기대값을 구하는 과정이다. 제곱합을 구하는 일반적인 방법은 분산분석에 의한 기법을 이용하고 있다. 본 논문에서는 행렬표현의 모형식을 이용하여 요인에 따른 제곱합을 모형행렬로 생성된 공간으로의 사영으로 계산하는 방법을 제안하고 있다. 제곱합의 기대값 또는 평균평방합의 기대값을 구하는 방법으로 가정된 분석모형을 이용한 대수적 방법이나 Hartley (1967)의 합성법을 적용할 수 있다. 그러나 Hartely의 방법은 기대값의 계산과정에 있어 많은 양의 계산을 요구하고 있어 이용하기가 쉽지 않은 점이 있다. 본 논문에서는 제곱합의 기대값 계산에 있어 분산성분의 계수로 주어지는 관련행렬의 대각합이 그 행렬의 고유근들의 합이라는 점에 착안하여 계산과정을 단순화하고 있다.

참고문헌

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

Choi, J. (2008b). A marginal probability model for repeated polytomous response data. *Journal of the Korean Data & Information Science Society*, **19**, 577-585.

Choi, J. (2010). A mixed model for repeated split-plot data. *Journal of the Korean Data & Information Science Society*, **21**, 1-9.

Galecki, A. T. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics-Theory and Methods*, **23**, 3105-3119.

Graybill, F. A. (1976). *Theory and application of the linear model*, Wadsworth Publishing Company, Inc., Belmont.

Graybill, F. A. (1983). *Matrices with applications in statistics*, Wadsworth Publishing Company, Inc., Belmont.

Hartley, H. O. (1967). Expectations, variances and covariances of ANOVA mean squares by "synthesis". *Biometrics*, **23**, 467-480.

Johnson, R. A. and Wichern, D. W. (1988). *Applied multivariate statistical analysis*, 2nd edition, Prentice Hall, Inc., Englewood Cliffs.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, 2nd edition, Chapman and Hall, London.

Milliken, G. A. and Johnson, D. E. (1984). *Analysis of messy data*, Van Nostrand Reinhold, New York.

Searle, S. R. (1971). *Linear models*, John Wiley and Sons, Inc., New York.

Journal of the Korean Data & Information Science Society
2011, **22**(3), 381–387

한국데이터정보과학회지

Variance components in one-factor random model by projections

Jaesung Choi[1]

[1]Department of Statistics, Keimyung University

Received 2 March 2011, revised 22 April 2011, accepted 27 April 2011

Abstract

This paper suggests a method for estimating components of variance in one-factor random model. Estimates of variance components are given by the method of moments. Sums of squares due to variance sources are obtained by projections. This paper also shows how to use eigenvalues for getting the coefficients of variance components in the expression of the expectations of the mean squares. The suggested method shows easier and faster than the method of Harley's synthesis.

Keywords: Eigenvalue, method of moments, projection, random model, variance components.

[1] Professor, Department of Statistics, Keimyung University, Daegu 704-701, Korea.
E-mail: jschoi@kmu.ac.kr