응용통계연구 (2011)
24(2), 373–381
DOI: 10.5351/KJAS.2011.24.2.373

사영에 의한 제1종 분석

최재성[1]

[1]계명대학교 통계학과

(2011년 1월 접수, 2011년 3월 채택)

요 약

본 논문은 실험자료에 대한 분석모형으로 이원 분산분석모형을 가정한다. 고정효과 모형의 가정하에 요인별 변동량을 구하기 위한 방법으로 제1종 분석을 다루고 있다. 모형의 순차적 적합에 따라 얻어지는 요인별 제곱합의 계산방법으로 대수적 방법이 아닌 사영에 의한 분석방법을 제공한다. 관측자료를 다차원상의 공간벡터로 간주할 때, 최소제곱법에 의한 요인별 변동량은 계획행렬로 생성되는 모수추정 공간에서 요인별 부분공간으로의 사영에 이르는 거리제곱으로 구해질 수 있음을 논의하고 있다. 또한 사영행렬로 부터의 고유벡터와 고유근을 이용하여 요인별 변동량을 구하는 방법을 제공하고 있다. 균형자료나 불균형자료에서 모형의 순차적 적합에 따른 제1종 분석이 행해질 때 요인별 변동량의 합은 처리제곱합과 일치하나 제2종 분석의 경우 불균형자료에서 이러한 성질이 만족되지 않음을 논의하고 있다.

주요용어: 사영, 제1종 제곱합, 사영행렬, 균형자료, 고유벡터, 고유근.

1. 서론

실험에서 고려되는 처리들의 효과에 대한 추론은 일반적으로 선형모형의 가정하에 행해진다. 이때, 자료분석에 이용되는 선형모형은 처리구조와 실험단위들의 특성을 고려한 설계구조에 따라 다양한 유형의 분석모형으로 주어지게 된다. 실험단위 또는 개체의 반응에 영향을 미치는 요인들의 수와 실험의 특성은 행해지는 실험의 처리구조를 결정하게 되고, 반응을 나타내는 개체의 동질적 또는 이질적 특성을 나타내는 변수의 수에 따라 실험의 설계구조가 정해진다. Milliken과 Johnson (1984) 그리고 Steel과 Torrie (1980)는 처리구조와 설계구조에 따른 다양한 유형의 자료분석 모형의 가정하에 분석방법들을 다루고 있다. 특히, 모형내 확률효과와 관련된 분산성분들의 추론방법에 관한 논의는 Corbeil과 Searle (1976), Graybill (1976), Montgomery (1976) 그리고 Searle 등 (1992)의 문헌에서 살펴볼 수 있다. 최재성 (2008)은 처리구조내에 반복측정 요인이 포함될 때 모형과 자료분석 방법을 논의하고 있다. 처리들의 효과를 알아보기 위한 선형모형으로 평균모형(means model)이나 효과모형(effects model)을 생각할 수 있다. 평균모형이 효과모형에 비해 상대적으로 계산이 용이하고 해석의 편리한 이점이 있으나 거의 모든 통계 프로그램이 효과모형을 다루기 위해 개발되어 있으므로 실험자들은 오히려 효과모형을 선호하고 있다. 효과모형의 가정하에 자료분석이 행해질 때, 어떤 유형의 통계적 가설이 검정되는가를 유의해야 한다. 왜냐하면, 효과모형에서 다루어지는 모수들의 수는 자료로 부터 추정가능한 모수들의 수보다 많은 모수들을 포함하고 있기 때문에 추정가능한 모수들의 함수 (estimable functions of parameters)인 가를 확인해야 한다.

본 연구는 2010년도 계명대학교 비사연구기금으로 이루어졌음.
[1](704-701) 대구광역시 달서구 신당동 1000번지, 계명대학교 자연과학대학 통계학과, 교수.
E-mail: jschoi@kmu.ac.kr

본 논문에서는 균형자료(balance data) 또는 불균형자료(unbalanced data)를 분석하기 위한 이원 ANOVA 모형을 고려한다. 균형자료와 불균형자료에 대한 논의는 Milliken과 Johnson (1984) 그리고 Searle 등 (1992)에서 살펴볼 수 있다. 가정된 모형에서 변동요인에 따른 제곱합을 구하기 위한 방법으로 사영(projection)을 이용하는 방법을 제공하고자 한다. 사영에 의한 자료분석은 가정된 모형의 행렬표현으로 인해 자료의 표현이 간편하고 대수식에 의한 제곱합보다 계산의 용이함이 뛰어나다. 또한 기하학적인 측면에서 자료분석의 이해를 도울 수 있다. 사영을 이용할 때, 모형행렬과 연관된 고유벡타와 고유근의 확인을 통해 처리효과와 관련된 축의 변동량과 자유도의 계산이 간편하다. 이러한 관점은 다양한 유형의 고정효과모형의 자료분석에 쉽게 확대 적용되는 이점이 있다.

2. 이원 분산분석모형의 가정

사영을 이용한 제1종 제곱합(Type I Sum of Squares)을 구하는 방법을 다루기 위해 자료에 대한 가정을 살펴보기로 한다. 실험단위 또는 개체의 반응을 변수 Y라 둔다. 개체의 반응에 영향을 미치는 독립변수로 두 개의 요인 A와 B를 생각한다. A는 $i = 1, 2, \ldots, a$개의 수준을 갖고 B는 $j = 1, 2, \ldots, b$개의 고정된 수준들을 갖는 고정요인으로 가정한다. 따라서, 실험에서 고려되는 처리들의 수는 모두 ab개 이다. 요인 A의 수준 i에서의 수준효과를 α_i라 두면, a개의 α_i들은 고정효과들 이다. 요인 B의 수준 j에서의 수준효과를 β_j라 두면, b개의 β_j들도 고정효과를 나타낸다. 두 요인 A와 B의 교호작용을 δ_{ij}로 나타낸다. ab개의 처리들이 k개 실험단위에 임의로 배정되고 실험단위들은 동질적이라 간주한다. 이러한 가정하에 행해진 실험에서 수집된 자료를 분석하기 위한 선형모형식은 두 개의 유형으로 표현될 수 있다. 식으로 표현하기 위해 요인 A의 수준 i, 요인 B의 수준 j와의 수준결합인 처리 (i, j)가 실험단위 k에 행해졌을 때, 나타나는 반응을 y_{ijk}라 두자. 단, $k = 1, 2, \ldots, n_{ij}$이다. 처리 (i, j)가 행해진 실험단위들의 모집단에서 평균반응을 μ_{ij}라 두면 평균모형(means model)은 다음과 같다.

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}. \tag{2.1}$$

처리모집단의 평균이 독립변수들의 함수이고 관련모수들의 선형식으로 가정될 때, 선형적인 효과모형(effects model)으로 표현된다.

$$y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijk}. \tag{2.2}$$

단, μ는 실험단위들의 모집단에서 개체반응에 대한 평균을 나타낸다. 평균모형에 비해 효과모형을 선호하는 이유는 단순히 처리평균을 이용한 비교분석보다 처리구조에 따른 요인들의 효과를 파악하고, 요인간의 교호작용을 포함한 효과들의 구조적 성격을 규명하는 모형구축에 있다고 볼 수 있다. 또한 거의 모든 상용프로그램이 효과모형을 이용하고 있는 점도 간과할 수 없다. 벡터공간에서 사영을 위한 관측벡터 $\boldsymbol{y}$에 대한 행렬표현은 다음과 같다.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{\epsilon}, \tag{2.3}$$

여기서 $\boldsymbol{X}$는 식 (2.2)의 고정효과 모형내 모수들과 관련된 계수벡타들의 행렬을 나타내는 모형행렬이다. 모형행렬 $\boldsymbol{X}$의 크기는 $n \times p$이다. 여기서 n은 ab개 처리내 관측수들의 합으로 $\sum_{i,j} abn_{ij}$를 나타낸다. 벡터 $\boldsymbol{\tau}$는 $p \times 1$인 모수벡터이다. p는 모형내 모수들의 수이며 그 합은 $(1 + a + b + ab)$개 이다. $\boldsymbol{\epsilon}$은 크기가 $n \times 1$인 오차벡타이다. 식 (2.2)에서 식 (2.3)으로의 행렬표현은 일차원상의 관측자료를 다차원상의 관측벡터로 간주한다는 점이다. 각 처리내 관측자료의 변동이 동일하다는 가정하에서 분석이 이루어질 때, 자료분석의 일반적인 기법은 최소제곱법을 이용하여 식 (2.2)의 모수 추정 및 요인별 변동량을 계

산하게 된다. 일차원상의 관측자료를 다차원상의 관측벡터로 간주할 때, 최소제곱법에 의한 모수벡터의 추정 및 요인별 변동량의 계산은 다차원상의 관측벡터를 모형행렬 $\boldsymbol{X}$에 의해 생성된 부분공간으로의 사영에 의한 분석을 의미한다. 사영에 의한 분석을 위해 식 (2.3)의 모형행렬 $\boldsymbol{X}$를 μ와 각 요인별 효과를 나타내는 행렬과 모수벡터로 표현하면 다음과 같다.

$$\boldsymbol{y} = \boldsymbol{j}\mu + \boldsymbol{X_A}\boldsymbol{\alpha} + \boldsymbol{X_B}\boldsymbol{\beta} + \boldsymbol{X_{AB}}\boldsymbol{\delta} + \boldsymbol{\epsilon}. \tag{2.4}$$

단, μ는 전체평균을 나타내는 모수이다. μ의 계수벡터 $\boldsymbol{j}$는 n개의 원소가 모두 1인 열벡터이다. 다차원상의 한 공간벡터 $\boldsymbol{y}$에 대한 일반적인 가정은 $\text{MVN}(\boldsymbol{\mu}, \Sigma)$로 주어진다. 처리 (i,j)의 실험모집단에 대한 분포로 각기 평균 μ_{ij}이고 분산 σ^2인 정규분포를 가정할 때, 크기 n인 실험자료를 나타내는 벡터 $\boldsymbol{y}$는 $\text{MVN}(\boldsymbol{X\tau}, \sigma^2\boldsymbol{I})$인 분포로 부터의 관측벡터로 간주된다. 단, $\boldsymbol{X}$=$(\boldsymbol{j}, \boldsymbol{X_A}, \boldsymbol{X_B}, \boldsymbol{X_{AB}})$이고, $\boldsymbol{\tau}'$=$(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta})$임을 나타낸다.

관측벡터 $\boldsymbol{y}$에 대한 행렬표현식 (2.4)의 가정하에 요인별 변동량을 최소제곱법으로 구한다고 하자. 최소제곱법에 의한 처리제곱합은 모형의 적합방식에 따라 요인별 제곱합의 총합과 다를 수 있다. 최소제곱법에 의한 요인별 변동량을 구하는 방법의 하나는 모형을 순차적으로 적합시켜 제곱합을 구하는 순차적 분석방법을 적용할 수 있다. 이때 얻어지는 제곱합을 제1종의 제곱합이라 부른다.

다차원 공간상의 관측벡터로 간주하여 사영을 이용하는 경우에 모형의 순차적 적합이 아닌 방식으로 각 요인에 따른 제곱합이 구해질 때, 순차적 분석으로 얻어진 제곱합과 동일하지 않다. 오차제곱합을 최소화하는 방법의 최소제곱법이나 사영에 의한 분석방법은 동일한 자료에 대해 동일한 분석결과를 예상할 수 있다. 그러나 사영에 의한 분석방법이 일반적으로 이용되는 분산분석 방법보다 쉽고 효율적으로 계산할 수 있는 이점이 있다. 또한, 부분공간으로의 사영에 이용되는 사영행렬은 계수 1인 고유근과 고유벡터로 구성되는 행렬의 합으로 표현되므로 이를 활용한 분석기법들이 개발될 수 있다. 모형의 순차적 적합에 따른 제1종 분석은 처리내 자료의 양에 상관없이 처리제곱합과 일치하는 요인별 변동량을 구할 수 있다. 순차적 적합을 따르지 않는 모형비교의 다른 방법은 처리내 자료의 양이 같지 않을 때 처리제곱합은 요인별 제곱합과 일치하지 않음을 보여준다. 사영에 의한 분석은 이러한 제곱합의 차이에 대한 분석을 가능하게 해준다. 이러한 점에서 기존의 분산 분석방법을 개선하는 효과가 있다고 볼 수 있다. 여기서는 자료의 형태에서 균형자료와 불균형자료인 경우를 생각해 보기로 한다.

3. 사영을 이용한 제1종 제곱합

고정효과의 행렬모형식 (2.4)의 가정하에 제1종 분석을 이용한 제곱합을 구해본다. 자료의 분산분석에 적용되는 제1종 분석은 모형의 순차적 적합에 따른 잔차제곱합의 차로 제곱합을 구하고 있다. 잔차제곱합의 차를 이용한 요인별 변동량의 계산은 매번 잔차제곱합을 계산해야 하는 번거로움이 발생하게 된다.

그러나 사영행렬을 이용하는 경우에는 잔차제곱합의 계산없이 요인별 변동량을 구할 수 있다. 사영분석에 의한 방법을 살펴보기로 한다. n차원상의 관측벡터 $\boldsymbol{y}$의 제곱합을 TSS로 둘 때, TSS = $\boldsymbol{y'y}$로 계산된다. 순차적 적합에 따른 모수 μ의 적합량은 관측벡터 $\boldsymbol{y}$에 대한 모형식 (2.4)에서 다음 방정식을 자료에 적합시켜 얻게 된다.

$$\boldsymbol{y} = \boldsymbol{j}\mu + \boldsymbol{\epsilon_1}. \tag{3.1}$$

단, $\boldsymbol{y}$는 $n \times 1$인 열벡터, $\boldsymbol{j}$는 $n \times 1$인 계수 1의 열벡터 그리고 $\boldsymbol{\epsilon_1}$은 $n \times 1$인 오차벡터이다. 모수 μ의 계수벡터 $\boldsymbol{j}$에 의해 생성된 부분공간으로의 사영 $\boldsymbol{jj^-y}$을 나타내는 사영행렬 $\boldsymbol{jj^-}$은 $\boldsymbol{j(j'j)^{-1}j'}$이다. $\boldsymbol{j}$로의 사영에 따른 제곱거리를 SSJ라 둘 때, SSJ = $\boldsymbol{y'(jj^-)y}$가 된다. 모형내 포함된 모수의 순서대로 요인

A의 제곱합을 구하기 위한 방정식은

$$\boldsymbol{y} = \boldsymbol{j}\mu + \boldsymbol{X_A}\boldsymbol{\alpha} + \boldsymbol{\epsilon_2} \tag{3.2}$$

이다. 이 경우의 모형행렬을 $\boldsymbol{X_1}$이라 둘 때, $\boldsymbol{X_1}$=($\boldsymbol{j}, \boldsymbol{X_A}$)이다. 요인 A에 따른 제곱합을 구하기 위해 n차원상의 관측벡터 $\boldsymbol{y}$를 $\boldsymbol{X_1}$에 의해 생성된 벡터공간 $\mathcal{X}_1$으로의 사영을 구한다. 사영행렬 $\boldsymbol{X_1}\boldsymbol{X_1}^-$에 의한 사영은 $\boldsymbol{X_1}\boldsymbol{X_1}^-\boldsymbol{y}$로 정의된다. 요인 A의 변동량을 구하기 위한 사영행렬을 $\boldsymbol{X_{11}}\boldsymbol{X_{11}}^-$라 두자. $\boldsymbol{X_{11}}\boldsymbol{X_{11}}^-$=$\boldsymbol{X_1}\boldsymbol{X_1}^- - \boldsymbol{j}\boldsymbol{j}^-$로 정의된다. 요인 A의 제곱합을 SSA라 둘 때 SSA= $\boldsymbol{y'}\boldsymbol{X_{11}}\boldsymbol{X_{11}}^-\boldsymbol{y}$이다.

모형내 포함된 모수의 순서대로 요인 B의 제곱합을 구하기 위한 방정식은

$$\boldsymbol{y} = \boldsymbol{j}\mu + \boldsymbol{X_A}\boldsymbol{\alpha} + \boldsymbol{X_B}\boldsymbol{\beta} + \boldsymbol{\epsilon_3} \tag{3.3}$$

이다. 모형행렬을 $\boldsymbol{X_2}$라 둘 때, $\boldsymbol{X_2}$ = ($\boldsymbol{j}, \boldsymbol{X_A}, \boldsymbol{X_B}$)이다. 요인 B의 제곱합을 구하기 위한 사영행렬은 $\boldsymbol{X_2}\boldsymbol{X_2}^-$이다. $\boldsymbol{X_2}$에 의해 생성된 벡터공간 $\mathcal{X}_2$로의 사영은 $\boldsymbol{X_2}\boldsymbol{X_2}^-\boldsymbol{y}$이다. 요인 B의 변동량을 구하기 위한 사영행렬을 $\boldsymbol{X_{22}}\boldsymbol{X_{22}}^-$라 둘 때, $\boldsymbol{X_{22}}\boldsymbol{X_{22}}^- = \boldsymbol{X_2}\boldsymbol{X_2}^- - \boldsymbol{j}\boldsymbol{j}^- - \boldsymbol{X_{11}}\boldsymbol{X_{11}}^-$로 정의된다. 요인 B의 제곱합을 SSB라 둘 때, SSB = $\boldsymbol{y'}\boldsymbol{X_{22}}\boldsymbol{X_{22}}^-\boldsymbol{y}$이다.

모형내 포함된 교호작용 AB의 제곱합을 구하기 위한 방정식은

$$\boldsymbol{y} = \boldsymbol{j}\mu + \boldsymbol{X_A}\boldsymbol{\alpha} + \boldsymbol{X_B}\boldsymbol{\beta} + \boldsymbol{X_{AB}}\boldsymbol{\delta} + \boldsymbol{\epsilon_4} \tag{3.4}$$

이다. 이 경우의 모형행렬은 $\boldsymbol{X}$이다. 요인 AB의 제곱합을 구하기 위한 사영행렬은 $\boldsymbol{X}\boldsymbol{X}^-$이다. $\boldsymbol{X}$에 의해 생성된 벡터공간 $\mathcal{X}$로의 사영은 $\boldsymbol{X}\boldsymbol{X}^-\boldsymbol{y}$이다. 요인 AB의 변동량을 구하기 위한 사영행렬을 $\boldsymbol{X_{33}}\boldsymbol{X_{33}}^-$라 두자. $\boldsymbol{X_{33}}\boldsymbol{X_{33}}^- = \boldsymbol{X}\boldsymbol{X}^- - \boldsymbol{j}\boldsymbol{j}^- - \boldsymbol{X_{11}}\boldsymbol{X_{11}}^- - \boldsymbol{X_{22}}\boldsymbol{X_{22}}^-$로 정의된다. 요인 AB의 제곱합을 SSAB라 둘 때 SSAB = $\boldsymbol{y'}\boldsymbol{X_{33}}\boldsymbol{X_{33}}^-\boldsymbol{y}$로 구해진다. 고정효과모형의 가정에서 요인별 변동량을 구하기 위한 제1종 제곱합은 모형의 순차적 적합을 통해 구해지는 잔차제곱합의 차이를 이용하고 있다. 사영에 의한 분석은 제1종 분석에서와 같이 순차적인 모형의 적합을 통하여 얻어지는 사영행렬을 이용한다는 점에서 잔차제곱합을 이용하는 기존방법과는 차이가 있게 된다. 모형의 순차적 적합에서 주어지는 사영행렬을 이용한 요인별 제곱합에 대한 식은 다음과 같이 주어진다. 다음 식은 관측벡터 $\boldsymbol{y}$를 $\boldsymbol{X}$ = ($\boldsymbol{j}, \boldsymbol{X_A}, \boldsymbol{X_B}, \boldsymbol{X_{AB}}$)에 의해 생성된 벡터공간으로 상호직교하는 사영행렬에 의한 제곱합을 나타낸다.

$$\boldsymbol{y'}\boldsymbol{X}\boldsymbol{X}^-\boldsymbol{y} = \boldsymbol{y'}\boldsymbol{j}\boldsymbol{j}^-\boldsymbol{y} + \boldsymbol{y'}\boldsymbol{X_{11}}\boldsymbol{X_{11}}^-\boldsymbol{y} + \boldsymbol{y'}\boldsymbol{X_{22}}\boldsymbol{X_{22}}^-\boldsymbol{y} + \boldsymbol{y'}\boldsymbol{X_{33}}\boldsymbol{X_{33}}^-\boldsymbol{y}. \tag{3.5}$$

단, $\boldsymbol{j}$는 $n \times 1$인 열벡터를 나타낸다. 모형을 고려하지 않았을 때의 평균에 대한 총변동량은 $\sum_{i=1}^{n}(y_i - \bar{y})^2$이다. 사영에 의한 이 양은 $\boldsymbol{y'}(\boldsymbol{I} - \boldsymbol{j}\boldsymbol{j}^-)\boldsymbol{y}$이다. 제곱합 $\boldsymbol{y'}\boldsymbol{j}\boldsymbol{j}^-\boldsymbol{y}$은 $\boldsymbol{y}$를 평균 벡타인 $\boldsymbol{j}$로의 사영이 행해질 때, $\boldsymbol{j}\boldsymbol{j}^-\boldsymbol{y}$까지의 제곱거리를 나타낸다. 식 (3.5)에서 등호는 각 부분공간으로의 사영을 구하기 위한 사영행렬 $\boldsymbol{j}\boldsymbol{j}^-$, $\boldsymbol{X_{11}}\boldsymbol{X_{11}}^-$, $\boldsymbol{X_{22}}\boldsymbol{X_{22}}^-$ 그리고 $\boldsymbol{X_{33}}\boldsymbol{X_{33}}^-$이 모두 직교행렬일 때 성립한다.

요인이 둘인 이원 분산분석모형에서 고정효과의 기입순서대로 하나씩 순차적으로 모형을 적합시키는 과정에서 고정효과들의 변동량을 구하는 방법이 제1종 분석이다. 따라서, 계수가 r인 모형행렬 $\boldsymbol{X}$의 p개 열벡터로 생성된 벡터공간 $\mathcal{X}$는 r차원의 벡터공간이다. 제1종 분석은 r차원의 벡터공간 $\mathcal{X}$으로의 사영 $\boldsymbol{X}\boldsymbol{X}^-\boldsymbol{y}$를 순차적 적합에 따른 r개 고유벡터와 고유근으로 효과들의 변동량을 구하게 된다. 이때의 각 효과에 대한 변동량은 벡터공간 $\mathcal{X}$로의 사영행렬 $\boldsymbol{X}\boldsymbol{X}^-$이 r개의 고유벡터와 고유근으로 구성되는 행렬들의 합으로 표현될 때 계산되는 양이다. 즉,

$$\boldsymbol{X}\boldsymbol{X}^- = \lambda_1\boldsymbol{p_1}\boldsymbol{p_1'} + \lambda_2\boldsymbol{p_2}\boldsymbol{p_2'} + \cdots + \lambda_r\boldsymbol{p_r}\boldsymbol{p_r'} \tag{3.6}$$

이고

$$\boldsymbol{y'XX^{-}y} = \boldsymbol{y'}(\lambda_1 \boldsymbol{p_1 p_1'})\boldsymbol{y} + \boldsymbol{y'}(\lambda_2 \boldsymbol{p_2 p_2'})\boldsymbol{y} + \cdots + \boldsymbol{y'}(\lambda_r \boldsymbol{p_r p_r'})\boldsymbol{y} \tag{3.7}$$

이다. 여기서 λ_i는 고유벡터 $\boldsymbol{p_i}$ $(i = 1, 2, \ldots, r)$의 고유근을 나타낸다. 이 경우에 수집된 자료의 처리제곱합 $\boldsymbol{y'XX^{-}y}$은 요인별 제곱합의 총합과 일치함을 보여준다. 실험자료를 분석하기 위한 모형의 순차적 분석방법은 각 처리별 자료의 개수가 동일하던 동일하지 않던 처리제곱합은 요인별 제곱합의 총합과 일치하게 됨을 알 수 있다. 또한, 모형내 모수를 추정하기 위한 p개 열벡터는 모형의 순차적 적합으로 인해 벡터공간 $\mathcal{X}$내 r개 고유벡터로 치환된다. 이는 모형내 모수들은 더 이상 추정가능한 모수들이 아님을 의미하고 있다. 그러므로 고유벡터로의 사영은 추정가능한 모수들의 추정치를 제공하게 된다. 이때의 추정가능한 모수들이 제1종 추정가능모수로 정의된다.

처리에 따른 관측값의 개수가 동일하지 않을 때, 순차적 적합이 아닌 모형의 비교에 따른 분산분석법은 모형의 적합방식에 따라 모형에 따른 제곱합, 즉, 처리에 따른 제곱합은 변동요인별 제곱합과 일치하지 않음을 보여준다. 자료분석에 적합한 모형구축을 위한 모형비교의 한 방법인 제2종 분석(Type II Analysis)에서 계산된 변동요인별 제곱합은 제1종 분석과는 달리 처리제곱합은 요인별 변동제곱합의 합과 같지 않음을 보여준다.

제1종의 제곱합과 동일한 값을 얻기 위한 사영의 이용에서 또 하나의 방법은 계획행렬 $\boldsymbol{X}$의 변환을 생각할 수 있다. 계획행렬 $\boldsymbol{X}$의 열벡터 순서대로 순차적 사영을 적용한 Gram-Schmidt의 직교화 과정을 이용하여 직교변환된 행렬을 $\boldsymbol{Z}$라 둔다. 행렬 $\boldsymbol{Z}$에 의해 생성된 벡터공간으로의 사영을 이용할 경우에도 변동요인별 제곱합은 제1종 제곱합과 동일하게 된다. 왜냐하면, 계획행렬 $\boldsymbol{X}$에 의해 생성된 벡터공간 $\mathcal{X}$는 직교로 변환된 행렬 $\boldsymbol{Z}$에 의해 생성된 벡터공간 $\mathcal{Z}$와 동일공간이기 때문이다.

사영을 이용한 분석의 이점은 변동요인에 따른 제곱합의 계산이 기존의 분석방법과는 달리 간편하다는 점이다. 또한, 모형에서 고려된 독립변수간의 상관성에 구애됨이 없이 사영행렬 $\boldsymbol{XX^{-}}$을 이용하여 분석이 용이하다는 이점을 갖고 있다. 달리 말하면, 사영행렬의 고유값과 고유벡터를 활용하여 처리요인에 따른 자유도의 수와 해당축의 변동량을 계산해 낼 수 있는 장점을 갖고 있다. 모형의 순차적 적합방식에 따른 요인별 제곱합을 고유벡터와 고유근을 이용하여 구할 때, 해당하는 제곱합은 관측벡터 $\boldsymbol{y}$의 성분으로 주어지는 2차형식으로 표현된다. 제곱합의 계산은 이 2차형식과 관련된 사영행렬로 부터 주어지는 고유벡터와 고유근을 사용하게 된다.

4. 균형자료와 불균형자료에서의 제곱합

실험에서 주어진 자료가 균형자료(balanced data)인 경우에 제1종 분석이나 제2종 분석에 의한 요인들의 변동에 따른 제곱합의 총합은 처리에 따른 관측값들의 총제곱합과 일치하게 된다. 그러나 실험자료가 불균형자료(unbalanced data)일때, 제1종 분석에 따른 요인들의 변동량의 합은 처리에 따른 총변동량과 일치하나 제2종 분석에 의한 요인들의 변동량의 합은 처리에 따른 총변동량과 일반적으로 일치하지 않게 된다. 이러한 현상은 다른 유형의 제곱합에서도 관측된다. 여기서 제1종 분석과 제2종 분석을 비교하는 이유는 관심처치들의 효과를 예측하기 위한 모형구축에 관심을 둘 때 이용되는 분석방법이기 때문이다. 균형자료의 경우 관측벡타를 모형행렬로 생성되는 추정공간으로의 사영은 사영행렬 $\boldsymbol{XX^{-}}$에 의해 주어진다. 이 사영행렬은 요인들의 변동량을 구하기 위한 부분공간으로의 상호 직교하는 행렬들의 합으로 주어진다. 그러나 불균형자료의 경우 요인들의 변동과 관련된 부분공간들은 상호직교하지 않으므로 부분공간으로의 사영행렬들의 합은 모형행렬로 주어진 공간에서의 사영행렬 $\boldsymbol{XX^{-}}$과 일치하지 않게 된다. 왜냐하면 제1종 분석은 벡터공간 $\mathcal{X}$내에서 모형의 순차적 적합을 이용하므로 벡터

표 **5.1.** 불균형자료의 생성표

번호	요인 A	요인 B	관측값 y
1	a_1	b_1	2
2	a_1	b_2	6
3	a_1	b_3	8
4	a_2	b_1	6
5	a_2	b_2	3
6	a_2	b_3	14
7	a_3	b_1	12
8	a_3	b_2	9
9	a_3	b_3	6
10	a_3	b_3	5

공간 $\mathcal{X}$내 모든 부분공간은 상호직교하는 공간들로 구성되고 상호직교하는 공간으로의 사영은 좌표축을 나타내는 고유벡터로의 사영에 이르는 제곱거리에 해당하며 처리간의 관측수에 상관없이 일정값을 제공하게 된다. 그러나 제2종 분석은 모형내 추정하고자 하는 모수효과와 동일수준의 효과 또는 낮은 수준의 모든 효과에 적합시킨 후 부분공간으로의 사영을 이용하기 때문에 사영이 행해지는 공간은 상호 직교하지 않는 부분공간으로 구성될 수 있다. 사영이 행해진 공간이 직교하지 않는 부분공간으로 구성될 때 모수효과들의 추정에 중복계산되는 일정한 변동량이 생기게 된다. 균형자료의 경우 추정이 행해지는 벡터공간 $\mathcal{X}$는 직교하는 부분공간들로 구성되므로 부분공간들에 행해지는 사영간에 내적은 존재하지 않게 된다.

5. 사영에 의한 자료분석의 예

사영에 의한 실험자료를 분석하기 위해 실험단위의 반응 y에 영향을 미치는 두 요인 A와 B의 수준결합에서 행해진 실험을 가정한다. 요인 A는 a_1, a_2, a_3의 세 수준이고, 요인 B도 b_1, b_2, b_3의 세 수준이라 하자. 수준결합으로 주어지는 각 처리(a_i, b_j), $(i = 1, 2, 3,\ j = 1, 2, 3)$에서의 관측수가 동일하지 않은 다음의 불균형자료를 가정한다.

위 자료를 분석하기 위한 모형으로 식 (2.4)를 가정할 때, 모형행렬 $\boldsymbol{X}$에서 $\boldsymbol{X_A} = (\boldsymbol{X}_{a_1}, \boldsymbol{X}_{a_2}, \boldsymbol{X}_{a_3})$인 행렬이다. $\boldsymbol{X_B} = (\boldsymbol{X}_{b_1}, \boldsymbol{X}_{b_2}, \boldsymbol{X}_{b_3})$인 행렬이고, $\boldsymbol{X_{AB}} = (\boldsymbol{X}_{ab_{11}}, \boldsymbol{X}_{ab_{12}}, \ldots, \boldsymbol{X}_{ab_{33}})$인 행렬을 나타낸다. $\boldsymbol{X}_{a_i}$, $\boldsymbol{X}_{b_j}$ 그리고 $\boldsymbol{X}_{ab_{ij}}$는 원소가 0과 1로 구성되는 열벡타를 표시한다. 제1종 분석에 의한 제곱합을 나타내기 위해 관측변수 y에 따른 변동량을 TSS라 둘 때, TSS $= \boldsymbol{y'y}$이고, $\boldsymbol{y'y} = 631$이다. 모형행렬 $\boldsymbol{X}$로의 사영에 의한 제곱거리를 SSM이라 둘 때, SSM $= \boldsymbol{y'X}\boldsymbol{X}^{-}\boldsymbol{y}$이고 계산값은 504.1이다. 오차공간으로의 사영은 $(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^{-})\boldsymbol{y}$이고 제곱거리를 SSE라 둘 때, SSE $= \boldsymbol{y'}(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^{-})\boldsymbol{y}$가 된다. 계산값은 0.5이다. 처리에 따른 제곱합을 모형의 순차적 적합에 따른 평균 μ와 요인별 변동량을 SSJ, SSA, SSB, SSAB로 나타내기로 한다. 전평균 μ에 따른 변동량 SSJ $= \boldsymbol{y'}(\boldsymbol{jj}^{-})\boldsymbol{y}$이고 그 값은 504.1이다. 사영행렬 $\boldsymbol{jj}^{-}$의 고유근과 고유벡터로도 제곱합을 계산할 수 있다. 사영행렬 $\boldsymbol{jj}^{-}$는 한개의 고유근과 고유벡터로 표현된다. 고유근을 λ_1라 둘 때, λ_1은 1이고 고유벡터를 $\boldsymbol{p_1}$이라 둘 때, $\boldsymbol{p_1} = (-0.3162278) \times \boldsymbol{j}$이다. 따라서, 고유근과 고유베터를 이용한 제곱합은 $\boldsymbol{y'}(\lambda_1(\boldsymbol{p_1}\boldsymbol{p_1'}))\boldsymbol{y}$에 의해 계산되고 그 값은 504.1로 동일하게 얻어진다. 요인 A에 따른 변동량 SSA는 식 (3.5)의 $\boldsymbol{y'}(\boldsymbol{X_{11}}\boldsymbol{X_{11}}^{-})\boldsymbol{y}$이고 그 값은 13.57이다. 평균벡터 $\boldsymbol{j}$로의 사영을 나타내는 사영행렬 $\boldsymbol{jj}^{-}$와 평균벡터에 적합된 요인 A로의 사영행렬 $\boldsymbol{X_1}\boldsymbol{X_1}^{-}$는 상호 직교하는 행렬이다.

즉, $(\boldsymbol{jj}^{-})(\boldsymbol{X_{11}}\boldsymbol{X_{11}}^{-})$는 모든 원소가 0인 7×7행렬이다. 요인 B의 처리벡터로 구성되는 행렬 $\boldsymbol{X_B}$로

의 사영에 따른 요인 B의 제곱합, SSB = 7.52381이다. 교호작용을 나타내는 행렬 $\boldsymbol{X_{AB}}$에 의해 생성된 부분공간으로의 사영에 의한 요인 AB의 제곱합 SSAB= 105.3095이다. 두 요인들의 수준결합으로 주어지는 6개 처리(a_i, b_j) $(i = 1, 2,\ j = 1, 2, 3)$에서의 제곱합을 SST라 둘 때, $\boldsymbol{y'(XX^- - jj^-)y}$로 계산된다. 처리제곱합 SST = SSA + SSB + SSAB = 126.4000이다. SST를 계산하기 위한 사영행렬은 모형행렬 $\boldsymbol{X}$의 적합에서 얻어진 행렬 $\boldsymbol{XX^-}$에서 $\boldsymbol{jj^-}$를 제외한 행렬을 이용한다. 즉, 사영행렬을 $\boldsymbol{X_T X_T{}^-}$라 둘 때, $\boldsymbol{X_T X_T{}^- = XX^- - jj^-}$이고 사영까지의 제곱거리는 $\boldsymbol{y'(X_T X_T{}^-)y}$이다. 이 제곱거리는 고유근이 1인 8개 고유벡터로의 사영에 이르는 제곱거리의 합으로도 계산된다. 관측벡타 $\boldsymbol{y}$의 성분들로 주어지는 2차형식의 제곱합에 나타나는 사영행렬들은 상호직교하는 대칭의 멱등행렬로 주어진다.

실험으로부터 얻어지는 자료의 유형이 각 처리의 관측수가 동일한 균형자료의 경우에는 제1종 분석 또는 제2종 분석에 따른 처리평방합은 요인별 평방합의 합과 일치하나 불균형자료인 경우에는 제2종 분석에 따른 처리평방합은 요인별 평방합의 합과 일치하지 않음을 나타낸다. 주어진 자료에 대해 제2종 분석을 적용하면, TSS, SSJ, SSB, SSAB는 동일하나 SSA는 제1종 분석에서의 값과는 다른 값을 나타내게 되고 따라서, SST의 값도 달라진다. 이러한 현상은 모형의 적합방식에 따라 사영이 행해지는 부분공간의 고유벡터들이 모형행렬 $\boldsymbol{X}$로 생성된 벡터공간의 고유벡터들로 구성되지 않음을 의미한다. 따라서 상호 직교하지 않는 부분공간으로의 사영에 따른 요인별 변동량은 모형적합 방식에 따라 다른 값을 나타내게 된다. 균형자료의 경우 모형행렬로 생성된 모수추정의 벡터공간은 상호직교하는 부분공간들로 구성되므로 모형의 적합방식에 상관없이 요인별 변동량은 일정하게 된다. 제2종 분석에 따른 SSA를 구하기 위해 $\boldsymbol{X_b}$를 $\boldsymbol{X_b = (XX^- - jj^-)X_B}$라 둔다. 이 때, $\boldsymbol{X_b X_b{}^-}$는 모수 추정공간에서 행렬 $\boldsymbol{X_B}$로의 사영행렬을 나타낸다. 모수 추정공간에서 이 사영행렬을 제외한 요인 A의 변동량을 구하기 위한 행렬을 $\boldsymbol{X_{t|b}}$라 두자. $\boldsymbol{X_{t|b} = (XX^- - jj^- - X_b X_b{}^-)X_A}$이다. 제2종 분석에 따른 요인 A의 변동량은 사영행렬 $\boldsymbol{X_{t|b} X_{t|b}{}^-}$에 의해 계산되는 $\boldsymbol{y' X_{t|b} X_{t|b}{}^- y}$이고 그 값은 11.60714로 주어진다.

6. 결론

본 논문은 실험자료의 분석을 위한 이원 분산분석모형으로 고정효과 모형을 가정하고 있다. 실험단위의 반응에 영향을 미치는 요인별 변동량의 계산은 일변량의 관측자료가 다변량의 관측벡터로 간주될 때, 다차원상의 관측벡터가 모형행렬로 생성되는 모수추정공간의 사영이라는 점에서 부분 벡터공간으로의 사영행렬을 이용한 요인들의 제곱합으로 구할 수 있음을 논의하고 있다. 따라서, 고정효과모형의 자료분석방법으로 사영에 의한 분석방법을 제기하고 있다. 사영을 이용한 자료의 분석기법은 대수식에 의한 일반적인 분산분석방법보다 효율적인 분석방법을 제공할 뿐만 아니라 분산분석의 기본적인 원리에 추가된 다양한 분석기법이 논의될 수 있음을 보여준다. 사영에 의한 제1종 분석에서 모형의 순차적 적합에 따라 주어지는 사영행렬은 요인별 제곱합을 나타내는 2차형식의 행렬임을 보여주고 있다. 또한, 각 제곱합에 관련된 사영행렬들은 상호 직교하는 대칭의 멱등행렬임을 나타내고 있다. 모형의 순차적 적합에 의한 제1종 분석은 처리내 자료개수에 종속되지 않는 일정 양으로 계산됨을 상호 직교하는 대칭의 멱등행렬로 설명하고 있다. 반면에 모형의 순차적 적합을 따르지 않는 제2종 분석은 처리내 자료의 양이 달라질 때, 상호 직교하지 않는 사영행렬에 의해 계산됨을 논의하고 있다.

반응벡타 $\boldsymbol{y}$가 다차원 공간상의 벡타로 간주되므로 $\boldsymbol{y}$에 대한 분포로 다변량 정규분포를 가정한다. 오차제곱합을 최소로 하는 최소제곱법은 관측벡타 $\boldsymbol{y}$를 계획행렬 $\boldsymbol{X}$에 의해 생성된 벡타공간 $\mathcal{X}$로의 사영임을 의미한다. 계획행렬로 생성되는 벡타공간으로의 사영을 이용한 순차적 분석을 행할 때, 분산분석에서 행해지는 계산방법과는 달리 사영행렬로 변동요인에 따른 제곱합을 간편하게 계산할 수 있는 분석방법을 다루고 있다. 사영을 이용한 순차적 분석의 자료분석 방법은 다양한 분산분석모형에 이용되어 질

수 있는 이점을 갖고 있다. 또한, 제1종 제곱합의 계산에 있어 사영행렬의 고유근과 고유벡타를 이용할 수 있음을 보여주고 있다.

참고문헌

최재성 (2008). 반복측정의 다가 반응자료에 대한 일반화된 주변 로짓모형, <응용통계연구>, **32**, 621–630.

Corbeil, R. R. and Searle, S. R. (1976). A comparison of variance component estimators, *Biometrics*, **32**, 779–791.

Graybill, F. A. (1976). *Theory and Application of the Linear Model*, Wadsworth, California.

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

Montgomery, D. C. (1976). *Design and Analysis of Experiments*, John Wiley & Sons, New York.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components*, John Wiley & Sons, New York.

Steel, R. G. and Torrie, J. H. (1980). *Principles and Procedures of Statistics*, McGraw-Hill, New York.

Type I Analysis by Projections

Jaesung Choi[1]

[1]Department of Statistics, Keimyung University

(Received January 2011; accepted March 2011)

Abstract

This paper discusses how to get the sums of squares due to treatment factors when Type I Analysis is used by projections for the analysis of data under the assumption of a two-way ANOVA model. The suggested method does not need to calculate the residual sums of squares for the calculation of sums of squares. Therefore, the calculation is easier and faster than classical ANOVA methods. It also discusses how eigenvectors and eigenvalues of the projection matrices can be used to get the calculation of sums of squares. An example is given to illustrate the calculation procedure by projections for unbalanced data.

Keywords: Projection, Type I Analysis, unbalanced data, projection matrix, eigenvector, eigenvalue.

The present research has been conducted by the Bisa Research Grant of Keimyung University in 2010.

[1]Professor, Department of Statistics, Keimyung University, 1000 Shindang-Dong, Dalseo-Gu, Daegu 704-701, Korea. E-mail: jschoi@kmu.ac.kr