*The Korean Journal of Applied Statistics* (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.7.1155
**29**(7), 1155–1163

# Mixed-effects model by projections

Jaesung Choi[a,1]

[a]Department of statistics, Keimyung University

(Received March 2, 2016; Revised May 13, 2016; Accepted July 7, 2016)

**Abstract**

This paper deals with an estimation procedure of variance components in a mixed effects model by projections. Projections are used to obtain sums of squares instead of using reductions in sums of squares due to fitting both the assumed model and sub-models in the fitting constants method. A projection matrix can be obtained for the residual model at each step by a stepwise procedure to test the hypotheses. A weighted least squares method is used for the estimation of fixed effects. Satterthwaite's approximation is done for the confidence intervals for variance components.

Keywords: fitting constants method, mixed-effects, projection matrices, stepwise procedure, weighted least squares method

## 1. 서론

실험단위의 반응에 영향을 주는 요인들로 고정요인과 확률요인이 포함되어 있을 때 실험자료를 분석하기 위한 모형으로 혼합효과의 선형모형을 가정하게 된다. 혼합모형의 가정에서 행해지는 분석은 고정효과의 분석과 확률효과의 분석으로 이루어진다. 혼합모형의 분석에 관한 연구는 Milliken과 Johnson (1984) 그리고 Searle (1971) 등의 많은 문헌에서 다루어지고 있다. 혼합모형의 분석방법으로 적률법, 최대우도법, MINQUE 방법 등을 이용할 수 있다.

실험자료의 분석모형으로 혼합모형을 가정할 수 있을 때 실험단위의 반응벡터를 $\boldsymbol{y}$라 두면 $\boldsymbol{y}$의 벡터공간에서 사영에 의한 분석이 가능하다. Choi (2011, 2012)는 자료분석을 위한 다양한 모형의 가정하에 벡터공간에서 정의되는 사영의 관점에서 분석하는 방법을 논의하고 있다. 사영에 의한 혼합모형의 분석을 위해 기존의 자료분석 방법 중 적률법이 이용될 때 적률법에 의한 분석과정에서 사영이 어떻게 활용될 수 있는가를 논의하고 그 결과가 동일함을 입증해 보이고자 한다. 이는 사영에 의한 분석이 자료분석의 다양한 선형모형하에서 효율적으로 이용될 수 있음을 보여주며 자료분석의 또 다른 측면에서 접근할 수 있는 방법을 제공하게 된다.

혼합모형의 가정하에 적률법으로 자료분석을 하는 경우에 확률성분의 추론을 위해 일반적으로 상수적합법(fitting constants method) 또는 Henderson (1953) 방법 III(Henderson's Method III)에 의해서 변동요인들의 제곱합을 구하게 된다. 그러나 고정효과 부분의 추론에는 다양한 모형비교방식을 이

This research was supported by the Keimyung-Scholar Research Grant of Keimyung University in 2016.
[1]Department of Statistics, Keimyung University, 1095 Dalgubeoldaero, Dalseogu, Daegu 42601, Korea.
E-mail: jschoi@kmu.ac.kr

용할 수 있게 된다. 분산성분의 추정을 위한 제곱합의 계산에 이용되는 상수적합법은 제곱합에서의 감소(reduction in sum of squares)로 주어지고 $\mathrm{R}(\cdot\,|\,\cdot)$로 표시된다. 사영에 의한 제곱합의 계산도 동일하게 구해짐을 다루게 된다. 혼합효과의 고정효과 부분에 대한 분석에도 사영이 어떻게 활용되는가를 논의하게 된다.

본 논문은 혼합모형을 이루는 두 유형의 효과를 추론하기 위한 분석과정에 벡터공간에서 정의되는 사영을 자료분석에 활용하는 방법을 제시하고 사영과 관련된 성질들을 이용한 자료분석의 효율성을 논의하는 데 초점을 두고 있다. 사영과 관련된 자세한 논의는 Johnson과 Wichern (1988) 그리고 Graybill (1983) 등에서 볼 수 있다.

## 2. 모형의 가정

실험자료의 분석을 위한 일반적인 혼합모형의 행렬표현식은 다음과 같다.

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\delta}_2 + \boldsymbol{X}_3\boldsymbol{\delta}_3 + \cdots + \boldsymbol{X}_k\boldsymbol{\delta}_k + \boldsymbol{\epsilon}. \tag{2.1}$$

단, $\boldsymbol{y}$는 $n \times 1$인 관측벡터이고 $\boldsymbol{X}_1$은 원소가 0 또는 1로 구성되며 크기가 $n \times p$인 고정효과벡터의 계수행렬로 계수(rank)가 $q\,(< p)$이다. $\boldsymbol{\beta}_1$은 $p \times 1$인 모수벡터이다. $\boldsymbol{X}_i$, $i = 2, 3, \ldots, k$는 확률효과벡터 $\boldsymbol{\delta}_i$의 계수행렬로 0 또는 1로 구성되며 크기가 $n \times c_i$인 완전계수행렬(full rank matrix)이다. $\boldsymbol{\delta}_i$는 $N(\boldsymbol{0}, \sigma_i^2\boldsymbol{I}_{c_i})$인 분포를 따르며 $\boldsymbol{\delta}_i$, $i = 2, 3, \ldots, k$는 상호 독립이라고 가정한다. $\boldsymbol{\epsilon}$은 $n \times 1$인 오차벡터이며 $N(\boldsymbol{0}, \sigma_\epsilon^2\boldsymbol{I}_n)$인 분포를 따른다고 가정한다. 오차벡터와 확률효과벡터들은 서로 독립으로 가정한다. 식 (2.1)은 $\boldsymbol{X}_1\boldsymbol{\beta}_1$으로 주어지는 고정효과부분과 $\boldsymbol{X}_2\boldsymbol{\delta}_2 + \cdots + \boldsymbol{X}_k\boldsymbol{\delta}_k + \boldsymbol{\epsilon}$인 확률효과부분의 두 성분으로 구성되므로 분석도 고정성분과 확률성분의 두 부분으로 나누어 행해진다. 혼합모형의 전체적 분석은 확률성분의 분석후에 고정효과의 분석이 가능하므로 확률성분의 분석을 위해

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1 \tag{2.2}$$

인 고정효과모형으로 표현한다. 단, $\boldsymbol{\epsilon}_1 = \boldsymbol{X}_2\boldsymbol{\delta}_2 + \cdots + \boldsymbol{X}_k\boldsymbol{\delta}_k + \boldsymbol{\epsilon}$이다. $\mathrm{Var}(\boldsymbol{\epsilon}_1) = \boldsymbol{\Sigma}$라 두면 $\boldsymbol{\Sigma}$는

$$\boldsymbol{\Sigma} = \sigma_2^2\boldsymbol{X}_2\boldsymbol{X}_2' + \sigma_3^2\boldsymbol{X}_3\boldsymbol{X}_3' + \cdots + \sigma_k^2\boldsymbol{X}_k\boldsymbol{X}_k' + \sigma_\epsilon^2\boldsymbol{I}_n \tag{2.3}$$

이다. 고정효과모형으로 변환된 식 (2.2)을 최소제곱법을 이용하여 자료에 적합시켜 잔차를 구한다. 잔차를 $\boldsymbol{r}_1$이라 두자. Moore-Penrose의 일반화된 역행렬을 이용한 정규방정식의 해벡터 $\hat{\boldsymbol{\beta}}_1$은 $\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1^-\boldsymbol{y}$로 구해지므로

$$\begin{aligned}\boldsymbol{r}_1 &= (\boldsymbol{I} - \boldsymbol{X}_1\boldsymbol{X}_1^-)\boldsymbol{y} \\ &= (\boldsymbol{I} - \boldsymbol{X}_1\boldsymbol{X}_1^-)\boldsymbol{X}_2\boldsymbol{\delta}_2 + \cdots + (\boldsymbol{I} - \boldsymbol{X}_1\boldsymbol{X}_1^-)\boldsymbol{X}_k\boldsymbol{\delta}_k + (\boldsymbol{I} - \boldsymbol{X}_1\boldsymbol{X}_1^-)\boldsymbol{\epsilon}\end{aligned} \tag{2.4}$$

이다. 식 (2.4)의 잔차모형은 혼합모형의 고정효과부분인 $\boldsymbol{X}_1\boldsymbol{\beta}_1$에 종속되지 않는 확률모형이다. 상수적합법으로 불리우는 Henderson 방법 III은 잔차의 확률모형에서 분산성분을 구하는 방법을 제공하나 본 논문에서는 사영을 이용하여 분산성분을 구하는 방법을 살펴보기로 한다. 분산성분에 관련된 논의는 Searle 등 (1992)에서 보여진다.

## 3. 잔차의 확률모형에 대한 사영

잔차모형에 대해 사영을 정의해 보기로 한다. 식 (2.4)의 잔차확률모형으로부터 잔차벡터 $\boldsymbol{r}_1$은 $(n - q)$차원의 벡터부분공간에 속하는 $n$개 성분의 열벡터이다. 잔차벡터가 $k$개 성분벡터들의 합으로 구성되

므로 $\boldsymbol{r}_1$의 벡터공간을 $k$개 분산성분별 벡터공간으로 분할하는 방법을 이용한다. 분할하는 방법으로 모형의 적합방식 중 제1종 제곱합을 구하는 단계별 방법(stepwise procedure)을 가정한다. 단계별 방법에 의한 $\boldsymbol{\delta}_2$의 계수행렬$(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^-)\boldsymbol{X}_2$로의 사영을 구하기 위한 모형은

$$\boldsymbol{r}_1 = (\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^-)\boldsymbol{X}_2\boldsymbol{\delta}_2 + \boldsymbol{\epsilon}_2 \tag{3.1}$$

이다. 확률벡터 $\boldsymbol{\delta}_2$의 추정을 위한 계수행렬 $(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^-)\boldsymbol{X}_2$로의 사영은 $[(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^-)\boldsymbol{X}_2][(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^-)\boldsymbol{X}_2]^-\boldsymbol{r}_1$이다. $\boldsymbol{r}_1$에서 $(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^-)\boldsymbol{X}_2$로의 사영을 뺀 잔차를 $\boldsymbol{r}_2$라 두면

$$\begin{aligned}\boldsymbol{r}_2 &= \boldsymbol{r}_1 - [(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^-)\boldsymbol{X}_2][(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^-)\boldsymbol{X}_2]^-\boldsymbol{r}_1 \\ &= (\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^- - [(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^-)\boldsymbol{X}_2][(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^-)\boldsymbol{X}_2]^-)\boldsymbol{y} \\ &= (\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^- - \boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-)\boldsymbol{y}\end{aligned} \tag{3.2}$$

로 표현된다. 단, $\boldsymbol{X}_{2p} = (\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^-)\boldsymbol{X}_2$이다. 식 (3.2)에서 $(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^- - \boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-)$는 $\boldsymbol{X}_1$과 $\boldsymbol{X}_2$와의 곱이 영행렬이므로 $\boldsymbol{\delta}_3$에 대한 모형은

$$\boldsymbol{r}_2 = \left(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^- - \boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-\right)\boldsymbol{X}_3\boldsymbol{\delta}_3 + \boldsymbol{\epsilon}_3 \tag{3.3}$$

이다. $\boldsymbol{\delta}_3$를 추정하기 위한 공간으로의 사영은 $[(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^- - \boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-)\boldsymbol{X}_3][(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^- - \boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-)\boldsymbol{X}_3]^-\boldsymbol{r}_2$이다. $\boldsymbol{X}_{3p}$를 $[(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^- - \boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-)\boldsymbol{X}_3]$라 두면 $\boldsymbol{X}_{3p}$로의 사영행렬은 $\boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^-$이다. 잔차벡터를 $\boldsymbol{r}_3$라 두면

$$\begin{aligned}\boldsymbol{r}_3 &= \boldsymbol{r}_2 - \boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^-\boldsymbol{r}_2 \\ &= \left(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^- - \boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^- - \boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^-\left(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^- - \boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-\right)\right)\boldsymbol{y}\end{aligned} \tag{3.4}$$

이다. $\boldsymbol{\delta}_4$에 대한 모형은

$$\boldsymbol{r}_3 = \left(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^- - \boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^- - \boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^-\right)\boldsymbol{X}_4\boldsymbol{\delta}_4 + \boldsymbol{\epsilon}_4 \tag{3.5}$$

이다. $\boldsymbol{\delta}_4$를 추정하기 위한 공간으로의 사영은 $[(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^- - \boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^- - \boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^-)\boldsymbol{X}_4][(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^- - \boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^- - \boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^-)\boldsymbol{X}_4]^-\boldsymbol{r}_3$이다. $\boldsymbol{X}_{4p}$를 $[(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^- - \boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^- - \boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^-)\boldsymbol{X}_4]$라 두면 $\boldsymbol{X}_{4p}$로의 사영행렬은 $\boldsymbol{X}_{4p}\boldsymbol{X}_{4p}^-$이다. $\boldsymbol{\delta}_k$를 추정하기 위한 사영행렬은 단계별 방법에 의한 잔차벡터 $\boldsymbol{r}_{k-1}$의 모형으로부터 구해진다. 모형은

$$\boldsymbol{r}_{k-1} = \left(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^- - \boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^- - \boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^- - \cdots - \boldsymbol{X}_{(k-1)p}\boldsymbol{X}_{(k-1)p}^-\right)\boldsymbol{X}_k\boldsymbol{\delta}_k + \boldsymbol{\epsilon}_k \tag{3.6}$$

이다. $\boldsymbol{X}_{kp}$를 $\boldsymbol{\delta}_k$의 추정과 관련된 계수행렬이라 두자. $\boldsymbol{X}_{kp} = (\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^- - \boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^- - \boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^- - \cdots - \boldsymbol{X}_{(k-1)p}\boldsymbol{X}_{(k-1)p}^-)\boldsymbol{X}_k$로 주어지고 $\boldsymbol{X}_{kp}$로의 사영행렬은 $\boldsymbol{X}_{kp}\boldsymbol{X}_{kp}^-$로 표시된다. 확률벡터의 분산성분 추정에 필요한 제곱합은 단계별 방법으로 주어지는 부분공간으로의 사영을 아용할 수 있다. 단계별 방법에 의한 부분공간으로의 사영을 나타내는 사영행렬들은 직교한다. 따라서 고정효과부분이 제외된 확률효과와 관련된 이차형식의 제곱합은

$$\boldsymbol{y}'\left(\boldsymbol{I}-\boldsymbol{X}_1\boldsymbol{X}_1^-\right)\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-\boldsymbol{y} + \boldsymbol{y}'\boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^-\boldsymbol{y} + \cdots + \boldsymbol{y}'\boldsymbol{X}_{kp}\boldsymbol{X}_{kp}^-\boldsymbol{y} \tag{3.7}$$

로 분할된다. 식 (3.7)은 상호직교하는 부분공간으로의 사영과 관련된 사영행렬의 이차형식을 나타내고 있다. 이는 Henderson의 방법 III라 불리우는 상수적합법의 적용에서 제곱합에서의 감소를 나타내는

$R(\ )$대신에 단계별 방법의 적용으로부터 사영행렬을 이용하여 분산성분과 관련된 제곱합을 구하는 방법을 제공하고 있다. 식 (3.7)에서 $\boldsymbol{y}'\boldsymbol{X}_{ip}\boldsymbol{X}_{ip}^{-}\boldsymbol{y}$, $i=2,3,\ldots,k$의 기댓값은

$$\begin{aligned} E\left(\boldsymbol{y}'\boldsymbol{X}_{ip}\boldsymbol{X}_{ip}^{-}\boldsymbol{y}\right) &= \mathrm{tr}\left(\boldsymbol{X}_{ip}\boldsymbol{X}_{ip}^{-}\boldsymbol{\Sigma}\right)+\boldsymbol{\beta}_1'\boldsymbol{X}_1'\boldsymbol{X}_{ip}\boldsymbol{X}_{ip}^{-}\boldsymbol{X}_1\boldsymbol{\beta}_1 \\ &= \mathrm{tr}\left(\boldsymbol{X}_{ip}\boldsymbol{X}_{ip}^{-}\boldsymbol{\Sigma}\right) \end{aligned} \tag{3.8}$$

이다. 식 (3.8)에서

$$\begin{aligned} \mathrm{tr}\left(\boldsymbol{X}_{ip}\boldsymbol{X}_{ip}^{-}\boldsymbol{\Sigma}\right) &= \mathrm{tr}\left[\left(\boldsymbol{X}_{ip}\boldsymbol{X}_{ip}^{-}\right)\left(\sigma_2^2\boldsymbol{X}_2\boldsymbol{X}_{\boldsymbol{2}}'+\sigma_3^2\boldsymbol{X}_3\boldsymbol{X}_{\boldsymbol{3}}'+\cdots+\sigma_k^2\boldsymbol{X}_k\boldsymbol{X}_{\boldsymbol{k}}'+\boldsymbol{\epsilon}\right)\right] \\ &= \sum_{u=2}^{k}\sigma_i^2\mathrm{tr}\left[\left(\boldsymbol{X}_{\boldsymbol{u}}'\boldsymbol{X}_{ip}\boldsymbol{X}_{ip}^{-}\boldsymbol{X}_{\boldsymbol{u}}\right)\right]+\sigma_\epsilon^2\mathrm{tr}\left(\boldsymbol{X}_{ip}\boldsymbol{X}_{ip}^{-}\right) \end{aligned} \tag{3.9}$$

이다. 분산성분들의 벡터를 $\boldsymbol{\sigma}_\delta^2$라 두면 $\boldsymbol{\sigma}_\delta^2=(\sigma_2^2,\sigma_3^2,\ldots,\sigma_k^2,\sigma_\epsilon^2)'$이다. $\boldsymbol{y}$의 이차형식들로 구성되는 벡터를 $\boldsymbol{w}$로 두면 $\boldsymbol{w}=(\boldsymbol{y}'\boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^{-}\boldsymbol{y},\boldsymbol{y}'\boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^{-}\boldsymbol{y},\ldots,\boldsymbol{y}'\boldsymbol{X}_{kp}\boldsymbol{X}_{kp}^{-}\boldsymbol{y},\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{X}^{-})\boldsymbol{y})'$이다. 단, $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_k)$이다. 각 이차형식의 기댓값을 나타내는 방정식으로부터 구해지는 $\boldsymbol{\sigma}_\delta^2$의 계수행렬을 $\boldsymbol{A}$라 두면 다음의 선형방정식계

$$A\boldsymbol{\sigma}_\delta^2=\boldsymbol{w} \tag{3.10}$$

로 부터 분산성분의 추정치인 해를 얻게 된다.

## 4. 자료분석의 예

다음은 어떤 부품의 제조에 이용되는 기계의 교체를 위한 실험자료이다. 자료는 세 종류의 기계 1, 2, 3 중 하나를 선정하기 위한 실험으로부터의 생산성점수를 나타낸다. 실험은 회사의 자체직원 중 임의로 선정된 6명이 각 기계를 세번 작동하여 평가한 점수로 주어진 Milliken과 Johnson (1984)의 자료이다.

혼합모형의 사례연구(case study)로 제공된 실험자료를 나타내는 Table 4.1의 분석모형을 생각해 보자. 효능의 비교에 이용되는 기계는 세종류의 1, 2, 3으로 고정되어 있으므로 1, 2, 3은 고정요인의 세 수준을나타낸다. 이들 효과를 각기 $\beta_i$, $i=1,2,3$이라 두면 $\beta_i$, $i=1,2,3$은 고정효과이다. 실험에 참여하는 직원은 회사직원 중 임의로 선정되므로 확률요인이고 이들이 기계의 생산성에 미치는 효과는 확률효과로 간주된다. 확률효과를 $\delta_j$, $j=1,2,\ldots,6$으로 두면 $\delta_j$는 $N(0,\sigma_\delta^2)$를 따르는 서로 독립인 확률변수로 가정된다. 고정효과와 확률효과의 교호작용은 확률효과이므로 교호작용을 $\gamma_k$, $k=1,2,\ldots,18$로 두면 $\gamma_k$는 $N(0,\sigma_\gamma^2)$을 따르는 서로 독립인 확률변수로 가정된다. 모형의 행렬표현식은 식 (2.1)로부터

$$\boldsymbol{y}=\boldsymbol{j}\mu+\boldsymbol{X}_1\boldsymbol{\beta}+\boldsymbol{X}_2\boldsymbol{\delta}+\boldsymbol{X}_3\boldsymbol{\gamma}+\boldsymbol{\epsilon} \tag{4.1}$$

로 표현된다. 단, $\boldsymbol{j}$는 모든 원소가 1인 $54\times1$인 열벡터이고 $\mu$는 전체평균을 나타낸다. $\boldsymbol{X}_1$은 고정효과벡터 $\boldsymbol{\beta}$의 계수행렬로 크기가 $54\times3$인 불완전계수의 행렬이다. $\boldsymbol{\beta}$는 크기가 $3\times1$인 열벡터이다. $\boldsymbol{X}_2$는 확률벡터 $\boldsymbol{\delta}$의 계수행렬로 크기가 $54\times6$이다. $\boldsymbol{\delta}$는 크기가 $6\times1$인 열벡터이며 $N(\boldsymbol{0},\sigma_\delta^2\boldsymbol{I}_6)$인 분포를 따른다고 가정한다. $\boldsymbol{X}_3$는 확률벡터 $\boldsymbol{\gamma}$의 계수행렬로 크기가 $54\times18$이다. $\boldsymbol{\gamma}$는 크기가 $18\times1$인 열벡터이며 $N(\boldsymbol{0},\sigma_\gamma^2\boldsymbol{I}_{18})$인 분포를 따른다고 가정한다. $\boldsymbol{\epsilon}$에 대한 가정은 $E(\epsilon)=\boldsymbol{0}$이고 $\mathrm{var}(\boldsymbol{\epsilon})=\sigma^2\boldsymbol{I}_{54}$인 정규분포를 따른다고 가정한다. 확률성분을 추정하기 위해 식 (4.1)의 고정효과부분을 적합시킨 잔차벡터를 $\boldsymbol{r}_1$으로 두고 $\boldsymbol{r}_1$을 구한다. 즉,

$$\boldsymbol{y}=\left(\boldsymbol{j},\ \boldsymbol{X}_1\right)\begin{pmatrix}\mu\\ \boldsymbol{\beta}\end{pmatrix}+\boldsymbol{\epsilon}_1 \tag{4.2}$$

**Table 4.1.** Productivity sores data for Machine-Person

| Machine | Person | Score | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| 1 | 1 | 52.0 | 52.8 | 53.1 |
| 1 | 2 | 51.8 | 52.8 | 53.1 |
| 1 | 3 | 60.0 | 60.2 | 58.4 |
| 1 | 4 | 51.1 | 52.3 | 50.3 |
| 1 | 5 | 50.9 | 51.8 | 51.4 |
| 1 | 6 | 46.4 | 44.8 | 49.2 |
| 2 | 1 | 62.1 | 62.6 | 64.0 |
| 2 | 2 | 59.7 | 60.0 | 59.0 |
| 2 | 3 | 68.6 | 65.8 | 69.7 |
| 2 | 4 | 63.2 | 62.8 | 62.2 |
| 2 | 5 | 64.8 | 65.0 | 65.4 |
| 2 | 6 | 43.7 | 44.2 | 43.0 |
| 3 | 1 | 67.5 | 67.2 | 66.9 |
| 3 | 2 | 61.5 | 61.7 | 62.3 |
| 3 | 3 | 70.8 | 70.6 | 71.0 |
| 3 | 4 | 64.1 | 66.2 | 64.0 |
| 2 | 5 | 72.1 | 72.0 | 71.1 |
| 2 | 6 | 62.0 | 61.4 | 60.5 |

의 적합으로부터 $\boldsymbol{r}_1$을 구한다. $(\boldsymbol{j}, \boldsymbol{X}_1)$으로의 사영을 $(\boldsymbol{j}, \boldsymbol{X}_1)(\boldsymbol{j}, \boldsymbol{X}_1)^-$로 두면

$$\boldsymbol{r}_1 = (\boldsymbol{I} - (\boldsymbol{j}, \boldsymbol{X}_1)(\boldsymbol{j}, \boldsymbol{X}_1)^-)\boldsymbol{y} \tag{4.3}$$

로 구해진다. 잔차벡터 $\boldsymbol{r}_1$에 단계별 방법을 적용하여 $\boldsymbol{\delta}$의 추정을 위한 모형으로

$$\boldsymbol{r}_1 = \boldsymbol{X}_2\boldsymbol{\delta} + \boldsymbol{\epsilon}_2 \tag{4.4}$$

를 가정한다. 여기서 $\boldsymbol{r}_1 = \boldsymbol{I} - \boldsymbol{j}\boldsymbol{j}^- - \boldsymbol{X}_1\boldsymbol{X}_1^-$로 구해진다. 식 (4.4)로부터 $\boldsymbol{X}_2$로의 사영행렬을 $\boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-$로 나타내면 $\boldsymbol{X}_{2p} = (\boldsymbol{I} - \boldsymbol{j}\boldsymbol{j}^- - \boldsymbol{X}_1\boldsymbol{X}_1^-)\boldsymbol{X}_2$이다. $\boldsymbol{y}'\boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-\boldsymbol{y} = 1241.89$이다. $\boldsymbol{r}_2 = (\boldsymbol{I} - \boldsymbol{j}\boldsymbol{j}^- - \boldsymbol{X}_1\boldsymbol{X}_1^- - \boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-)\boldsymbol{r}_1$으로 주어진다. 잔차벡터 $\boldsymbol{r}_2$를 이용하여 $\boldsymbol{\gamma}$의 추정을 위한 모형으로

$$\boldsymbol{r}_2 = \boldsymbol{X}_3\boldsymbol{\gamma} + \boldsymbol{\epsilon}_3 \tag{4.5}$$

를 가정한다. 모형으로부터 $\boldsymbol{X}_3$로의 사영을 얻기위한 사영행렬은 $\boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^-$로 표시되고 $\boldsymbol{X}_{3p} = (\boldsymbol{I} - \boldsymbol{j}\boldsymbol{j}^- - \boldsymbol{X}_1\boldsymbol{X}_1^- - \boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-)\boldsymbol{X}_3$이다. $\boldsymbol{y}'\boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^-\boldsymbol{y} = 426.53$이다. 분산성분 ${\sigma_\epsilon}^2$과 관련된 제곱합의 계산은 $\boldsymbol{X} = (\boldsymbol{j}, \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3)$라 둘 때 $\boldsymbol{y}$를 $\boldsymbol{X}$로의 사영으로부터 구한다. $\boldsymbol{X}$로의 사영을 나타내는 사영행렬은 $\boldsymbol{X}\boldsymbol{X}^-$이므로 사영까지의 거리제곱합은 $\boldsymbol{y}'\boldsymbol{X}\boldsymbol{X}^-\boldsymbol{y}$이다. 따라서 $\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^-)\boldsymbol{y}$가 ${\sigma_\epsilon}^2$을 추정하기 위한 잔차제곱합이다. 자료로부터 계산된 $\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^-)\boldsymbol{y} = 33.29$이다. 분산성분들을 구하기 위한 선형방정식을 구성하기 위해 각 제곱합의 기댓값을 구한다.

$$\begin{aligned} E\left(\boldsymbol{y}'\boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-\boldsymbol{y}\right) &= \mathrm{tr}\left(\boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-\boldsymbol{\Sigma}\right) \\ &= \sigma_\delta^2 \mathrm{tr}\left[\boldsymbol{X}_2'\left(\boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-\right)\boldsymbol{X}_2\right] + \sigma_\gamma^2 \mathrm{tr}\left[\boldsymbol{X}_3'\left(\boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-\right)\boldsymbol{X}_3\right] + \sigma_\epsilon^2 \mathrm{tr}\left(\boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^-\right) \\ &= 45\sigma_\delta^2 + 15\sigma_\gamma^2 + 5\sigma_\epsilon^2 \end{aligned} \tag{4.6}$$

를 얻게 된다.

$$\begin{aligned} E\left(\boldsymbol{y}'\boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^{-}\boldsymbol{y}\right) &= \mathrm{tr}\left(\boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^{-}\boldsymbol{\Sigma}\right) \\ &= \sigma_\delta^2 \mathrm{tr}\left[\boldsymbol{X}_2'\left(\boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^{-}\right)\boldsymbol{X}_2\right] + \sigma_\gamma^2 \mathrm{tr}\left[\boldsymbol{X}_3'\left(\boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^{-}\right)\boldsymbol{X}_3\right] + \sigma_\epsilon^2 \mathrm{tr}\left(\boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^{-}\right) \\ &= 30\sigma_\gamma^2 + 10\sigma_\epsilon^2 \end{aligned} \tag{4.7}$$

를 얻게 된다.

$$\begin{aligned} E\left[\boldsymbol{y}'\left(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{X}^{-}\right)\boldsymbol{y}\right] &= \mathrm{tr}\left[\left(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{X}^{-}\right)\boldsymbol{\Sigma}\right] \\ &= \sigma_\delta^2 \mathrm{tr}\left[\boldsymbol{X}_2'\left(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{X}^{-}\right)\boldsymbol{X}_2\right] + \sigma_\gamma^2 \mathrm{tr}\left[\boldsymbol{X}_3'\left(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{X}^{-}\right)\boldsymbol{X}_3\right] + \sigma_\epsilon^2 \mathrm{tr}\left(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{X}^{-}\right) \\ &= 36\sigma_\epsilon^2 \end{aligned} \tag{4.8}$$

를 얻게 된다. 제곱합의 기댓값을 나타내는 식으로부터 분산성분을 얻기 위한 다음의 선형방정식계를 구성한다.

$$\boldsymbol{Q}\boldsymbol{\zeta} = \begin{pmatrix} 45 & 15 & 5 \\ 0 & 30 & 10 \\ 0 & 0 & 36 \end{pmatrix} \begin{pmatrix} \sigma_\delta^2 \\ \sigma_\gamma^2 \\ \sigma_\epsilon^2 \end{pmatrix} = \begin{pmatrix} 1241.89 \\ 426.53 \\ 33.29 \end{pmatrix}. \tag{4.9}$$

식 (4.9)로부터 해 벡터 $\hat{\boldsymbol{\zeta}} = (22.858, 13.909, 0.925)'$를 구할 수 있다. 제한가능도함수(REML)를 이용한 분산성분들의 추정벡터는 $(23.168, 17.891, 0.925)'$로 구해진다. 이는 추정방법에 따라 분산성분의 추정에 있어 다소의 차이가 있으나 상당히 유사한 결과를 보여주고 있음을 나타낸다. $\hat{\sigma}_\delta^2$에 대한 추정으로 식 (4.6)과 (4.7)을 이용할 때 $\hat{\sigma}_\delta^2 = (2/90)\boldsymbol{y}'(\boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^{-}\boldsymbol{y}) - (1/90)\boldsymbol{y}'(\boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^{-}\boldsymbol{y})$이다. Satterthwaite (1946)의 근사과정을 이용하여 $\chi^2$ 분포의 자유도 $r$을 구하면

$$\begin{aligned} r &= \frac{\left(\hat{\sigma}_\delta^2\right)^2}{\left(\frac{2}{90}\boldsymbol{y}'\left(\boldsymbol{X}_{2p}\boldsymbol{X}_{2p}^{-}\boldsymbol{y}\right)^2\right)/5 + \left(\frac{1}{90}\boldsymbol{y}'\left(\boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^{-}\boldsymbol{y}\right)^2\right)/10} \\ &= 3.38 \end{aligned} \tag{4.10}$$

으로 구해진다. 95% 신뢰구간을 구하기 위한 자유도 3.38에 해당하는 $\chi^2$값은 각기 ${\chi^2}_{(0.025,3.43)} = 0.307$과 ${\chi^2}_{(0.975,3.43)} = 10.046$으로 주어진다. 따라서, $\sigma_\delta^2$에 대한 95% 신뢰구간은

$$7.688 = \frac{3.38\hat{\sigma}_\delta^2}{{\chi^2}_{(0.975,3.43)}} < \sigma_\delta^2 < \frac{3.38\hat{\sigma}_\delta^2}{{\chi^2}_{(0.025,3.43)}} = 251.573 \tag{4.11}$$

으로 구해진다. $\hat{\sigma}_\gamma^2$에 대한 구간추정도 동일한 방법으로 구해진다. 식 (4.7)과 (4.8)로부터 $\hat{\sigma}_\gamma^2 = (36/10)\boldsymbol{y}'\boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^{-}\boldsymbol{y} - \boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{X}^{-})\boldsymbol{y}$이다. Satterthwaite의 근사과정을 이용하여 $\chi^2$ 분포의 자유도 $r'$을 구하면

$$\begin{aligned} r' &= \frac{\left(\hat{\sigma}_\gamma^2\right)^2}{\left(\frac{36}{10}\boldsymbol{y}'\boldsymbol{X}_{3p}\boldsymbol{X}_{3p}^{-}\boldsymbol{y}\right)/10 + \left(\boldsymbol{y}'\left(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{X}^{-}\right)\boldsymbol{y}\right)/36} \\ &= 1.13 \end{aligned} \tag{4.12}$$

로 구해진다. 95% 신뢰구간을 구하기 위한 자유도 1.13에 해당하는 $\chi^2$값은 각기 ${\chi^2}_{(0.025,1.13)} = 0.002$

와 $\chi^2{}_{(0.975,1.13)} = 5.370$으로 주어진다. 따라서, $\sigma_\gamma^2$에 대한 95% 신뢰구간은

$$2.927 = \frac{1.13\hat{\sigma}_\gamma^2}{\chi^2{}_{(0.975,1.13)}} < \sigma_\gamma^2 < \frac{1.13\hat{\sigma}_\gamma^2}{\chi^2{}_{(0.025,1.13)}} = 7858.585 \tag{4.13}$$

로 구해진다. 고정효과벡터$(\mu, \boldsymbol{\beta})' = (\mu, \beta_1, \beta_2, \beta_3)'$의 추정으로 가중최소제곱법을 이용하게 된다. 가중최소제곱법에 의한 추정벡터는 $(44.74, 7.62, 15.58, 21.53)$으로 구해진다. 모수벡터의 추정량에 대한 분산공분산행렬을 $\mathrm{cov}(\hat{\mu}, \hat{\boldsymbol{\beta}})'$라 두면,

$$\mathrm{cov}\begin{pmatrix}\hat{\mu}\\ \hat{\boldsymbol{\beta}}\end{pmatrix} = \begin{pmatrix} 2.5873 & 0.8624 & 0.8624 & 0.8624 \\ 0.8624 & 0.8624 & -0.5024 & -0.5024 \\ 0.8624 & -0.5024 & 1.8672 & -0.5024 \\ 0.8624 & -0.5024 & -0.5024 & 1.8672 \end{pmatrix} \tag{4.14}$$

로 구해진다. $\mu + \alpha_3$는 추정가능함수이고 추정값은 60.32로 주어진다. $\mathrm{var}(\hat{\mu} + \hat{\alpha_3}) = 6.1794$로 구해진다. 혼합효과모형은

$$y_{ijk} = \mu_i + \delta_j + \gamma_{ij} + \epsilon_{ijk} \tag{4.15}$$

로도 표현될 수 있다. 단, $\mu_i = \mu + \beta_i$이다. 이때의 행렬표현식은

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\mu} + \boldsymbol{X}_2\boldsymbol{\delta} + \boldsymbol{X}_3\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{4.16}$$

로 주어진다. 식 (4.15)의 평균벡터 $\boldsymbol{\mu}$의 추정벡터를 $\hat{\boldsymbol{\mu}}$이라 두면

$$\hat{\boldsymbol{\mu}} = \left(\boldsymbol{X}_1'\hat{\Sigma}^{-1}\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\hat{\Sigma}^{-1}\boldsymbol{y} \tag{4.17}$$

로 구해진다. $\hat{\boldsymbol{\mu}}$ 고정효과에 대한 추론방법은 자료가 균형인가 또는 불균형인가에 따라 다르다. 대부분의 균형혼합모형에서 $\boldsymbol{\mu}$의 추정벡터는 $\hat{\boldsymbol{\mu}} = (52.35, 60.32, 66.27)'$이다. 추정평균벡터의 분산공분산행렬을 $\mathrm{cov}(\hat{\boldsymbol{\mu}})$라 두면

$$\mathrm{cov}\,(\hat{\boldsymbol{\mu}}) = \begin{pmatrix} 6.1794 & 3.8097 & 3.8097 \\ 3.8097 & 6.1794 & 3.8097 \\ 3.8097 & 3.8097 & 6.1794 \end{pmatrix} \tag{4.18}$$

로 구해진다.

## 5. 결론

본 논문은 혼합효과모형의 가정하에서 분산성분의 추정과 고정효과추론에 사영이 어떻게 이용되는가를 논의하고 있다. 분산성분의 추정방법인 고정상수적합법(fitting constans method)에서 제곱합의 감소를 이용하는 방식 대신에 벡터공간에서 정의되는 사영을 활용하는 방법을 제시하고 있다. 사영에 의한 제곱합의 계산에 고정효과에 영향받지 않는 확률효과들로 구성되는 잔차모형을 제시하고 있으며 잔차모형에 단계별 방법(stepwise procedure)을 적용하여 얻어지는 모형행렬로의 사영을 통하여 제곱합을 구하는 방식을 제공하고 있다. 이 방법은 잔차제곱합에서의 감소를 이용하는 방법보다 효율적이며 벡터공간에서의 사영과 관련된 여러 개념들을 구체화하는 이점들이 있다.

또 다른 한편으로는 혼합모형에 상수적합법을 적용하여 유도된 잔차모형으로부터 분산성분을 추정하기 위한 모형행렬로의 사영과 사영행렬을 구하기 위해 단계별 적합방식을 논의하고 있다. 각 부분공간에서

계산된 제곱합의 기댓값을 이용하여 분산성분의 계수를 구하고 선형방정식계를 구성하는 방법을 제공하고 있다. 또한, 혼합모형에서 고정효과는 가중최소제곱법으로 모수추정벡터를 얻게 된다. 분산성분의 신뢰구간추정에서 해당하는 자유도를 구하기 위한 Satterthwaite의 근사적 과정이 설명되고 있다.

## References


Choi, J. S. (2011). Variance components in one-factor random model by projections, *Journal of the Korean Data & Information Science Society*, **22**, 381−387.

Choi, J. S. (2012). Type II analysis by projections, *Journal of the Korean Data & Information Science Society*, **23**, 155−1163.

Graybill, F. A. (1983). *Matrices with Applications in Statistics*, Wadsworth Publishing Company, Belmont.

Henderson, C. R. (1953). Estimation of variance and covariance components, *Biometrics*, **9**, 226−252.

Johnson, R. A. and Wichern, D. W. (1988). *Applied Multivariate Statistical Analysis* (2nd ed.), Prentice Hall, Englewood Cliffs.

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

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

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

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

# 사영에 의한 혼합효과모형

최재성[a,1]

[a]계명대학교 통계학과

(2016년 3월 2일 접수, 2016년 5월 13일 수정, 2016년 7월 7일 채택)

**요 약**

본 논문은 혼합효과의 선형모형에서 분산성분들의 추정방법으로 사영을 다루고 있다. 상수적합법에서 이용되는 제곱합에서의 감소(reductions in sums of squares) 대신에 사영을 이용하여 구하는 방법을 제시하고 있다. 단계별 방법에 의한 잔차모형으로부터 각 분산성분의 추정과 관련된 사영행렬을 구성하는 방법을 제공하고 있다. 사영행렬로 표현되는 이차형식의 기댓값을 이용하여 선형방정식계를 구성하고 적률법으로 분산성분을 추정하게 된다. 고정효과는 가중최소제곱법으로 추정되고 분산성분의 신뢰구간추정에 Satterthwaite의 근사과정으로 자유도를 계산하는 방법을 설명하고 있다.

주요용어: 단계별 방법, 사영행렬, 상수적합법, 혼합효과, 가중최소제곱법

이 연구는 2016년도 계명대학교 계명스칼라 연구기금으로 이루어졌음.

[1](42601) 대구광역시 달서구 딜구벌대로 1095, 계명대학교 통계학과. E-mail: jschoi@kmu.ac.kr