Journal of the Korean
Data & Information Science Society
1999, Vol. 10, No. 2, pp. 369 ~ 385

평균이동모형을 이용한 성장곡선모형의 이상점 진단에 관한 연구 [1]

심 규 박 [2]

요 약

성장곡선모형에서 다중 이상값들이나 영향관측값들을 탐지하는 문제는 선형회귀모형에서의 문제에 비해 매우 복잡하여 거의 이루어지지 않고 있는 실정이다. 본 연구에서는 이상점을 포함하고 있는 성장곡선모형에서 이들을 탐지하는 방법으로 평균이동모형을 이용하는 방법을 소개하였다. 이 방법을 이용하여 찾아낸 자료가 이상점인지의 여부를 예측표본재이용 의사 베이즈 우도 기준법을 이용한 등분산성의 검정을 통해 알아보았다. 끝으로 Potthoff(1964)등이 사용한 자료를 이용한 예제를 통해 이상점 탐지와 등분산성 검정을 실시한 결과를 제시였다.

주제어: 성장곡선모형, 이상점, 평균이동모형, 예측표본재이용 (Predictive Sample Reuse : PSR)

1. 서론

성장곡선모형(growth curve model) 은 분산분석에서 많이 사용하는 일반화된 다변량 모형 이다. 이 모형은 동물과 식물의 성장에 관한 모형으로 특히 유용하다. 이 모형은 Potthoff 등(1964)에 의해 처음 제안된 후 Geissor(1970), Rosen(1989, 1990) 및 Lee(1991)등 많은 학자들에 의해 연구되어 왔다. 성장곡선모형은

$$Y_{p\times n} = X_{p\times m}\ \tau_{m\times r}\ A_{r\times n}\ + \epsilon_{p\times n} \tag{1.1}$$

과 같은 형태를 지녔는데, 여기서 X와 A는 계수가 각각 $m < p$, $r < n$ 인 이미 알려진 계획행렬(design matrix)이고, 모수 τ는 모르는 수이다. 또한, 오차행렬 ϵ의 열들은 평균벡터 0이고, 미지의 합동공분산행렬 Σ를 가진 독립된 p차원 정규분포를 따른다. 즉,

[1]이 논문은 1999년도 동국대학교 학술연구비의 지원을 받아연구되었음
[2](780-714) 경북 경주시 석장동 707 동국대학교 자연과학대학 정보통계학과 조교수

$$G(Y|\tau,\ \Sigma) \ \sim \ N(X\tau A\ ,\ \Sigma \otimes I_n) \tag{1.2}$$

이며, 여기서, $\otimes$는 행렬의 Kronecker 곱이다.

Lee(1975)는 성장곡선을 베이지안의 관점에서 처음 연구하였는데, g개의 모집단 (Π_i, $i = 1,\ 2,\ \cdots,\ g$)으로부터 얻은 g개의 성장곡선이

$$G(Y_i|\tau_i,\ \Sigma_i,\ \Pi_i) \ \sim \ N(X\tau_i A_i\ ,\ \Sigma_i \otimes I_n) \tag{1.3}$$

인 분포를 따를 때, 미래 관측치행렬 $H_{p\times K}$의 분포는 아래 식을 따른다고 하였다.

$$G(H|\tau_i,\ \Sigma_i,\ \Pi_i) \ \sim \ \mathbf{N}(\mathbf{X}\tau_i\ \mathbf{F}_i,\ \mathbf{\Sigma}_i\ \otimes \mathbf{I}_K),\ i = 1,\ 2,\ \cdots,\ g \tag{1.4}$$

여기서, F_i는 A_i의 일부 열(column)들로 구성된 행렬이다.

일반적으로 성장곡선은 시간의 변화에 따른 다항식으로 나타나기 때문에 계획행렬 X는 다음과 같이 표시할 수 있다.

$$X \ = \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{m-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{m-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_p & t_p^2 & \cdots & t_p^{m-1} \end{bmatrix} \tag{1.5}$$

또한 행렬 A는 $p\times 1$차원 단위벡터(unit vector) E_i, $i = 1,\ 2,\ \cdots,\ r$와 영벡터(null vector)들의 조합으로 이루어 진다.

$$A^T_{r\times N} \ = \begin{bmatrix} E_1 & 0 & 0 & \cdots & 0 \\ 0 & E_2 & 0 & \cdots & 0 \\ 0 & 0 & E_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & E_r \end{bmatrix} \tag{1.6}$$

정규성의 가정 하에서 Rosen(1989)은 모수 τ와 Σ의 최우추정값(maximum likelihood estimate : MLE)들을 각각 아래와 같이 계산하였다.

$$\hat{\tau} \ = \ (X^T S^{-1} X)^{-1} X^T S^{-1} Y A^T (AA^T)^{-1} \tag{1.7}$$

$$\hat{\Sigma} \ = \ n^{-1}(Y \ - \ X\hat{\tau}A)(Y \ - \ X\hat{\tau}A)^T \tag{1.8}$$

여기서,

$$S = Y(I - A^T(AA^T)^{-1}A)Y^T \tag{1.9}$$

일반적인 선형모형 하에서 통계적 진단(statistical diagnostic) 문제에 대해서는 많은 학자들에 의해 회귀분석분야에서 이상점(outlier) 및 영향관측치(influential observation)를 탐지하는 주제로 연구되어 왔다. 그러나, 성장곡선모형에서 다중 이상값들이나 영향관측치들을 탐지하는 문제는 선형회귀모형에서의 문제에 비해 매우 복잡하여 거의 이루어지지 않고 있는 실정이다. 최근 Liski(1991)는 임의의 행렬 G가 양정치행렬이고, $\sigma^2 > 0$가 미지의 scalar일 때, 공분산행렬 $\Sigma = \sigma^2 G$를 가진 성장곡선모형에서 이상점들과 영향관측치들을 탐지하는 방법을 제안하였다. 또한 Pan등(1995)은 비구조화된 공분산행렬을 가진 성장곡선모형에서 이상점을 찾아내는 방법을 제안하였는데, 그는 다중개체제거모형 (Multiple-Individual -Deletion Model)등을 이용하였다.

성장곡선모형을 예로 들지는 않았으나 Hawkins등(1997)은 이상점을 가진 집단들의 선형판별분석을 행하는 과정에서 평균이동모형(mean-shift model)을 이용하여 이상점의 존재 여부를 판단한 바 있었다.

본 논문에서는 이들의 연구를 이용하여, 성장곡선모형에서 이상점이 존재하는 경우 이들을 찾아내어 제거한 후, 공분산행렬의 동일성검정을 실시하여 이 점의 포함여부에 따라 그룹간의 등분산성 검정에 긍정적인 영향을 미치는가를 비교하여 보았다.

2. 성장곡선모형의 모수 추정

$I = \{i_1, i_2, \cdots, i_k\}$ $(n > p + k)$를 제거해야 할 k개의 개체들에 대한 첨자를 포함한 집합이라 하자. 보편성의 상실이 없다면, 첨자집합들은 $I = \{n-k+1, n-k+2, \cdots, n\}$로 바꾸어 쓸 수 있으며, Y는 $Y = (Y_{(I)}, Y_I)$로 분할할 수 있다. 마찬가지로 행렬 A와 ϵ는 $A = (A_{(I)}, A_I)$ 및 $\epsilon = (\epsilon_{(I)}, \epsilon_I)$로 각각 분할 할 수 있다. 따라서, Y_I를 제거한 후의 성장곡선모형 (1.1)은 아래와 같이 쓸 수 있다.

$$Y_{(I)} = X\tau A_{(I)} + \epsilon_{(I)} \tag{2.1}$$

이 때, $\epsilon_{(I)} \sim N_{p,n-k}(0, \Sigma \otimes I_{n-k})$이다. 같은 방법으로 식(2.1)에 대한 τ와 Σ의 MLE는 각각

$$\hat{\tau}_{(I)} = (X^T S_{(I)}^{-1} X)^{-1} X^T S_{(I)}^{-1} Y_{(I)} A_{(I)}^T (A_{(I)} A_{(I)}^T)^{-1} \tag{2.2}$$

$$\hat{\Sigma}_{(I)} = (n-k)^{-1}(Y_{(I)} - X\hat{\tau}A_{(I)})(Y_{(I)} - X\hat{\tau}A_{(I)})^T \tag{2.3}$$

이다. 여기서,

$$S_{(I)} = Y_{(I)}(I_{(n-k)} - A_{(I)}^T(A_{(I)}A_{(I)}^T)^{-1}A_{(I)})Y_{(I)}^T. \tag{2.4}$$

Chatterjee 등(1988)은 $\hat{\tau}$와 $\hat{\tau}_{(I)}$ 사이의 관계를 아래와 같이 도출한 바 있다.

<정리 2-1> (Chatterjee(1988)). MLE $\hat{\tau}$와 $\hat{\tau}_{(I)}$와의 관계는 아래와 같다.

$$\hat{\tau}_{(I)} = \hat{\tau} - (X^TS^{-1}X)^{-1}X^TS^{-1}e_I\, V_I^{-1}K_I^{-1}\,(AA^T)^{-1} \tag{2.5}$$

여기서
$$V_I = I_k - H_I - e_I^TS^{-1}e_I + e_I^TS^{-1}X(X^TS^{-1}X)^{-1}X^TS^{-1}e_I \tag{2.6}$$
$$H_I = A_I^T(AA^T)^{-1}A_I$$
$$K_I = A_I - AY^TS^{-1}e_I + AY^TS^{-1}X(X^TS^{-1}X)^{-1}X^TS^{-1}e_I \tag{2.7}$$

이고, $e = (e_{(I)}, e_I)$ A상에 회귀된 Y의 잔차(residual)이다.

<정리 2-2> (Rosen(1990)). 다음 식

$$\begin{aligned} &S^{-1} - S^{-1}X(X^TS^{-1}X)^{-1}X^TS^{-1} \\ &= Q(Q^TSQ)^{-1}Q^T = S^{-1}Q_S = Q_S^TS^{-1} = Q_S^TS^{-1}Q_S \end{aligned} \tag{2.8}$$

을 사용하면 아래의 결과를 얻는다.

$$\begin{aligned} V_I &\equiv I_k - H_I - e_I^TQ_S^TS^{-1}Q_Se_I \\ K_I &\equiv A_I - AY^TQ_S^TQ_S^TS^{-1}Q_Se_I \end{aligned} \tag{2.9}$$

여기서, $Q \in \Xi$인데 Ξ는 행렬의 집합으로서 다음과 같이 정의된다.

$$\Xi = \left\{Q|Q : p \times (p-m),\ rank(Q) = p-m \ \text{ and } \ X^TQ = 0\right\}. \tag{2.10}$$

식 (2.8)과 (2.9)는 V_I와 K_I의 다른 형태이나, 두 식 모두 집합 Ξ에서 행렬 Q의 선택에 따라 달라지지 않는다.

Hawkins(1997)는 등분산성을 만족하는 판별분석에서 이상점을 탐지하는 방법에는 여러 가지가 있으나 평균이동모형을 보편적으로 사용할 수 있다고 하였다. 다음 모형을 생각해 보자.

$$Y_{p\times n} = X_{p\times m}\tau_{m\times r}A_{r\times n} + X_{p\times m}\psi_{m\times k}D_{k\times n} + \epsilon_{p\times n} \tag{2.11}$$

여기서, $\epsilon \sim N(0, \Sigma\otimes I_n)$이고, 새로운 모수 ψ는 평균이동모수, $D = (d_{n-k+1}, d_{n-k+2}, \cdots, d_n)^T$는 지시변수(indicator variable)의 행렬인데, D^T의 i번째열 d_i는 i번째 $(n-k+1 \le i \le n)$요소가 1이고 나머지는 0인 n차원 벡터이다. 따라서, $AD^T = A_I$, $YD^T = Y_I$이고 $DD^T = I_k$이 성립한다.

<정리 2-3> 평균이동모형에 대한 τ, ψ와 Σ의 MLE는 각각 다음과 같다.

$$\begin{aligned}
&\widehat{\tau}_a = \widehat{\tau}_{(I)} \\
&\widehat{\psi} = (X^T S_{(I)}^{-1} X)^{-1} X^T S_{(I)}^{-1} e_I (I_k - H_I)^{-1} \\
&\widehat{\Sigma}_a = n^{-1}\left\{(n-k)\widehat{\Sigma}_{(I)} + Q_{S_{(I)}} Y_I Y_I^T Q_{S_{(I)}}^T\right\}
\end{aligned} \tag{2.12}$$

<증명> $\widetilde{\tau} = (\tau, \psi), \widetilde{A}^T = (A^T, D^T)$라 두면, $Y \sim N(X\widetilde{\tau}\widetilde{A}, \Sigma \otimes I_n)$가 되고, $\widetilde{\tau}$와 Σ의 MLE는 각각 아래와 같다.

$$\widehat{\widetilde{\tau}} = (X^T S_a^{-1} X)^{-1} X^T S_a^{-1} Y \widetilde{A}^T (\widetilde{A}\widetilde{A}^T)^{-1} \tag{2.13}$$

$$\widehat{\Sigma}_a = n^{-1}\left\{S_a + Q_{S_a} Y P_{\widetilde{A}^T} Y^T Q_{S_a}^T\right\} \tag{2.14}$$

여기서, $P_{\widetilde{A}^T} = \widetilde{A}^T(\widetilde{A}^T\widetilde{A})\widetilde{A}$이고, $S_a = Y(I_n - P_{\widetilde{A}^T})Y^T$이다.

$\widehat{\tau}_a$와 $\widehat{\psi}$를 각각 평균이동모형에 대한 τ와 ψ의 MLE라 할 때, $\widehat{\widetilde{\tau}}$를 $\widehat{\widetilde{\tau}} = (\widehat{\tau}_a, \widehat{\psi})$로 분할하자. $P_{\widetilde{A}^T} = P_{A^T} + P_{(I_n - P_{A^T})D^T}$이므로 S_a에 대해 아래와 같이 유도할 수 있다.

$$\begin{aligned}
S_a &= Y(I_n - P_{\widetilde{A}^T})Y^T \\
&= Y(I_n - P_{A^T})Y^T - Y P_{(I_n - P_{A^T})D^T} Y^T \\
&= S - Y(I_n - P_{A^T})D^T\left\{D(I_n - P_{A^T})D^T\right\}^{-1} D(I_n - P_{A^T})Y^T \\
&= S - Y(I_n - P_{A^T})D^T(I_k - H_I)^{-1} D(I_n - P_{A^T})Y^T \\
&= S - e_I(I_k - H_I)^{-1} e_I^T
\end{aligned} \tag{2.15}$$

$$= S_{(I)}.$$

또한,

$$(\widetilde{A}\widetilde{A}^T)^{-1} = \begin{pmatrix} (\widetilde{A}_{(I)}\widetilde{A}_{(I)}^T)^{-1} & -(AA^T)^{-1}A_I(I_k - H_I)^{-1} \\ -\widetilde{A}_{(I)}^T(\widetilde{A}_{(I)}\widetilde{A}_{(I)}^T)^{-1} & (I_k - H_I)^{-1} \end{pmatrix}$$

이고,

$$Y\widetilde{A}^T(\widetilde{A}\widetilde{A}^T)^{-1} = (\widetilde{A}_{(I)}^T(\widetilde{A}_{(I)}\widetilde{A}_{(I)}^T)^{-1},\ e_I(I_k - H_I)^{-1})$$

이므로, 식 (2.13)에서 MLE $\widehat{\widehat{\tau}}$는 아래와 같다.

$$\begin{aligned} \widehat{\widehat{\tau}} &= (X^T S_{(I)}^{-1} X)^{-1} X^T S_{(I)}^{-1} \\ &\quad \cdot \left(Y_{(I)}\widetilde{A}_{(I)}^T(\widetilde{A}_{(I)}\widetilde{A}_{(I)}^T)^{-1},\ e_I(I_k - H_I)^{-1} \right) \\ &= \left(\widehat{\tau}_{(I)},\ (X^T S_{(I)}^{-1} X)^{-1} X^T S_{(I)}^{-1} e_I (I_k - H_I)^{-1} \right) \end{aligned} \tag{2.16}$$

인데, 정의에 의해 식 (2.16)은

$$\widehat{\tau}_a = \widehat{\tau}_{(I)}$$
$$\widehat{\psi} = (X^T S_{(I)}^{-1} X)^{-1} X^T S_{(I)}^{-1} e_I (I_k - H_I)^{-1}$$

를 의미한다.

식 (2.14)의 $\widehat{\Sigma}_a$와 식 (2.3)의 $\widehat{\Sigma}_{(I)}$사이의 관계를 도출하기 위하여, 식 (2.15)의 관계에서

$$Y P_{\widetilde{A}^T} Y^T = Y_{(I)} P_{\widetilde{A}_{(I)}^T} Y_{(I)}^T + Y_I Y_I^T \tag{2.17}$$

라 쓰면, 다음의 관계를 도출할 수 있다.

$$\begin{aligned} \widehat{\Sigma}_a &= n^{-1}(S_{(I)} + Q_{S_{(I)}} Y P_{\widetilde{A}^T} Y^T Q^T{}_{S_{(I)}}) \\ &= n^{-1}(S_{(I)} + Q_{S_{(I)}} Y_{(I)} P_{A_{(I)}^T} Y_{(I)}^T Q^T{}_{S_{(I)}} + Q_{S_{(I)}} Y_I Y_I^T Q^T{}_{S_{(I)}}) \\ &= n^{-1}\left\{ (n-k)\widehat{\Sigma}_{(I)} + Q_{S_{(I)}} Y_I Y_I^T Q_{S_{(I)}}^T \right\} \end{aligned} \tag{2.18}$$

정리 2-1과 2-3에서 MLEs $\widehat{\tau}$, $\widehat{\tau}_{(I)}$와 $\widehat{\tau}_a$사이의 관계를 알 수 있다. 비록, 위의 전개로 부터 Σ의 MLE들의 관계를 얻을 수 있다해도 $\widehat{\Sigma}_a$와 $\widehat{\Sigma}$사이의 관계도 명확히 알 수 있는 것은 아

니다. 그러나 Pan 등(1995)은 $\widehat{\Sigma}$ 행렬식에 대한 $\widehat{\Sigma}_a$ 행렬식의 비율을 계산하여 이상점 탐지에 사용한 바 있다. 그들이 제안한 $\widehat{\Sigma}_a$와 $\widehat{\Sigma}$사이의 비율은 아래와 같다.

$$\begin{aligned}\Lambda_I &\equiv det(\widehat{\Sigma}_a)/det(\widehat{\Sigma}) \\ &= det\left\{I_k - e_I^T S_{(I)}^{-1} X(X^T S_{(I)}^{-1} X)^{-1} X^T S_{(I)}^{-1} e_I (I_k - H_I + e_I^T S_{(I)}^{-1} e_I)^{-1}\right\}. \quad (2.19)\end{aligned}$$

혹은

$$\begin{aligned}T_I &\equiv det(\widehat{\Sigma})/det(\widehat{\Sigma}_a) \\ &= det\left\{I_k + e_I^T S^{-1} X(X^T S^{-1} X)^{-1} X^T S^{-1} e_I (I_k - H_I - e_I^T S^{-1} e_I)^{-1}\right\}. \quad (2.20)\end{aligned}$$

3. 이상점 진단

Hawkins 등(1997)은 평균이동모수 ψ가 0일 경우 평균이동모형은 이상점을 생성하지 않는다고 하였다. 따라서, 평균이동모수 ψ에 대해 아래와 같은 가설검정을 실시하여 보자.

$$H_0 : \psi = 0 \quad v.s. \quad H_a : \psi \neq 0. \quad (3.1)$$

귀무가설 H_0가 유의수준 α에서 기각되면, $Y_I = (y_{n-k+1},\ y_{n-k+2},\ \cdots,\ y_n)$는 유의수준 α 하에서 k개의 이상점들 이라 간주한다. Cook 등(1982)은 통계량 Λ_I가 $\Lambda_I \leq C_\alpha$이면 귀무가설 H_0은 유의수준 α하에서 기각됨을 보였다. 여기서, 통계량 Λ_I는 식(2.19)이고, C_α는 Wilk's 분포 $\Lambda(m,\ n-k-r-p+m,\ k)$의 $100\alpha\%$하측 임계점이다.

이 때, $k=1$이고 $I=\{i\}$ $(1 \leq i \leq n)$일 때, Wilk's 분포의 특성에 의해 귀무가설이 참이라는 가정 하에서 다음의 관계가 성립한다.

$$\frac{n-r-p}{m} \cdot \frac{1-\Lambda_i}{\Lambda_i} \sim F_{m,\ n-r-p} \quad (3.2)$$

그러므로, i번째 개체는 아래의 조건을 만족하면 하나의 이상점이 된다.

$$\begin{aligned}T_i &= 1 + \frac{e_i^T S^{-1} X(X^T S^{-1} X)^{-1} X^T S^{-1} e_i}{1 - p_{ii} - e_i^T S^{-1} e_i} \\ &\geq 1 + \frac{m}{(n-r-p)C_\alpha}, \quad (3.3)\end{aligned}$$

여기서, p_{ii}는 행렬 P_{A^T}의 i번째 대각요소이고, e_i는 잔차 e의 i번째 열이며, C_α는 분포 $\mathbf{F}_{m,\ n-r-p}$의 $100\alpha\%$하측 임계점이다.

4. PSR 성장곡선모형의 모수 추정

4.1 검정기준

관측치들의 모형 M_i, $i = 1, 2, \cdots, r$의 분포함수를 $\mathrm{F}(\cdot|M_i)$라 하면, 우도함수를 이용하여 최적의 모형을 선택할 수 있다. 그러나, 미지모수 θ_i들의 집합으로만 표본분포가 정의되는 모형 M_i가 있다고 할 때, Geisser(1979)등은 표본재이용의 개념을 이용하여 이러한 문제에 접근하였다. 이 방법은 관측치들의 결합조건부예측분포(joint conditional predictive distribution)들을 최대로 하는 모형을 선택하는 것으로서, PSR 의사 베이즈 우도기준법(Predictive Sample Reuse Quasi-Bayes Likelihood Criterion)이라 하였다. 즉, X를 모수 θ_i가 미지인 분포 $\mathrm{F}(\cdot|\theta_i, M_i)$, $\theta_i = (\mu_i, \Sigma_i)$를 가진 모형 M_i들 중 하나로부터 생성되었다고 가정했을 때,

$$X_{(j)} = [x_i, x_2, \cdots, x_{j-1}, x_{j+1}, \cdots, x_n] \tag{4.1}$$

을 j번째 관측벡터 x_j를 제외한 $p \times (n-1)$인 자료행렬이라 하자.

이 때, $\mathrm{F}(x|X, z, M_i)$는 모형 M_i를 사용하여 얻은 미래 관측치 y에 대한 베이지안 예측분포라 하면 베이지안 예측분포는 다음과 같이 쓸 수 있다.

$$\begin{aligned}&\mathrm{F}(x|X, z, M_i)\\ &\quad \propto \int \mathrm{F}(x|\theta_i, z, M_i) \prod_{i=1}^{n} \mathrm{F}(x_j|\theta_i, z_j, M_i) P(\theta_i) d\theta_i.\end{aligned} \tag{4.2}$$

여기서, $z_1, z_2, \cdots, z_n$은 대응관측치들이 속한 모집단이고, $P(\theta_i)$는 모형 M_i의 베이그 사전확률분포(vague prior)이다. 분포함수 $\mathrm{F}(x_j|X_{(j)},\ z_j,\ M_i)$를 관측치 x_j에 대한 베이지안 예측분포가 되도록 수정한 형태라면, x_j, $j = 1, 2, \cdots, n$의 베이지안 예측분포들의 곱을 구할 수 있다.

$$L_i = \prod_{j=1}^{n} \mathrm{F}(x_j|X_{(j)}, z_j, M_i), \quad i = 1, 2, \cdots, r \tag{4.3}$$

일 때, 식 (4.3)을 M_i의 PSR 의사 베이즈 우도라 한다. PSR 의사 베이즈 우도 기준법은 아래 식을 만족하는 L_i^*에 사용된 모형 M_i^*를 선택하는 것이다.

$$L_i^* = Max\{L_1, L_2, \cdots, L_r\} \tag{4.4}$$

4.2 다변량 정규분포에서의 등분산성 검정

$X_i = [x_{i1}, x_{i2}, \cdots, x_{in_i}]$를 p차원 다변량 정규분포함수를 가진 i번째 모집단으로 부터의 $p \times n_i$인 자료행렬이라 하자. 이 때, 공분산 행렬에 대한 r개의 가능한 모형을 아래와 같이 정의할 수 있다.

$$\begin{aligned} &M_1 : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_g \\ &M_2 : \Sigma_1 \neq \Sigma_2, \quad \Sigma_1 = \Sigma_3 = \cdots = \Sigma_g \\ &\qquad\vdots \\ &M_r : \Sigma_i \neq \Sigma_j \quad \text{for all } i \neq j \end{aligned} \tag{4.5}$$

여기서, $r = \sum_{m=1}^{g}\sum_{j=0}^{m}(-1)^{m-j}j^g/\{j!(m-j)!\}$이다.

이 때, $X = [x_1, x_2, \cdots, x_k]$에 대해 다음을 정의하자.

$$\begin{aligned} &X_{(ij)} : i\text{번째 자료행렬 } X_i\text{에 대해 } j\text{번째 관측벡터 } x_{ij}\text{가 빠진 자료} \\ &\bar{x}_i(j) = \sum_{s}^{(j)} x_{is}/(n_i - 1) \\ &S(k_t)^* = \frac{\sum_{u=a(t)u\neq i}^{b(t)}\left\{(n_u-1)S_u + \sum_{s}^{(j)}(x_{(is)} - X_i(j))(x_{(is)} - X_i(j))^T\right\}}{n(t) - k_t - 1} \\ &a(t) = \sum_{j=1}^{t-1} k_j + 1, \qquad b(t) = \sum_{j=1}^{t} k_j, \; t = 1, 2, \cdots, m \\ &n(t) = \sum_{i=a(t)}^{b(t)}, \qquad n = \sum_{i=i}^{K} n_i \\ &\text{단, } \sum_{s}^{(j)} \text{는} j\text{를제외한} S\text{의모든값들의합이다.} \end{aligned} \tag{4.6}$$

이 때, Geisser 등(1979)이 제안한 PSR 의사 베이즈 우도는 아래 식과 같이 정의할 수 있 다.

$$L(M^*) = \prod_{j=1}^{K} \mathrm{f}(x_{ij}|X(ij),\ i,\ M^*)$$

$$= \prod_{t=1}^{m} \prod_{i=a(t)}^{b(t)} \prod_{j=1}^{n_i} St_p\Big\{N(t) - k_t - 1,\ \overline{X}_i(j),\ n_i S(k_t)^*/(N_i - 1)\Big\} \tag{4.7}$$

여기서, $St_p(a,b,c)$는 p-차원 다변량 T-분포로서, 다음과 같이 정의된다.

$$St_p(a,b,c) = \frac{\Gamma\{(A+1)/2\}}{\pi^{p/2}\Gamma\{(a-p+1)/2\} \cdot |ac|^{1/2}\{1+(y-b)^T(ac)^{-1}(y-b)\}^{(a+1)/2}} \tag{4.8}$$

PSR 의사 베이즈 우도 기준법의 검정기준은 식 (4.4)에서 언급한 바와 같이 $L(M^*)$을 최대로 하는 모형 M^*을 선택하는 것이다.

4.3 성장곡선모형에서 PSR 의사 베이즈 우도 기준법을 이용한 등분산성 검정

Shim(1993)은 $g = 2$인 2개의 성장곡선모형의 경우 PSR 의사 베이즈 우도 기준을 적용하여 등분산성검정을 실시한 바 있다. 그는 2개의 다변량 정규모집단으로부터 생성된 2개의 성장곡선이 있을 때, 공분산 행렬에 대한 동일성 검정을 실시하기 위해 다음의 가설을 설정하였다.

$$\begin{aligned} H_0 &: \Sigma_1 = \Sigma_2 \\ H_a &: \Sigma_1 \neq \Sigma_2 \end{aligned} \tag{4.9}$$

위의 가설 하에서 모집단 모형을 각각

$$\begin{aligned} M_a &: \Sigma_1 = \Sigma_2 \\ M_b &: \Sigma_1 \neq \Sigma_2 \end{aligned} \tag{4.10}$$

라 하고, τ_i의 값은 미지이나 추정 가능한 값이라 하자. 첫째, $\Sigma_1 = \Sigma_2$의 경우 사전확률분포를 $h(\Sigma^{-1}) \propto |\Sigma|^{(p+1)/2}$이라 하면, M_a모형에서 다음과 같이 사후확률분포를 유도할 수 있다.

$$g(\Sigma^{-1}|Y) \propto |\Sigma|^{(n-p-1)/2} \exp\{-1/2 tr \Sigma^{-1}(Y - X\hat{\tau}A)(Y - X\hat{\tau}A)^T\}. \tag{4.11}$$

이 사후확률분포로부터 아래 식을 얻을 수 있다.

$$E(\Sigma|\tau, Y) = (n-p-1)^{-1}(Y - X\hat{\tau}A)(Y - X\hat{\tau}A)^T \tag{4.12}$$

이 때, 미래 관측치 y에 대한 예측분포는 다음과 같다.

$$G(y|\tau, Y, \Pi_i) = D(X\tau F_i, I, (Y - X\hat{\tau}A)(Y - X\hat{\tau}A)^T, n + K) \tag{4.13}$$

여기서, 임의의 행렬 $L_{p\times K}$에 대하여 확률밀도함수가 식 (4.13)과 같을 때 $D(\Delta,\ \Lambda,\ \Sigma,\ n)$은 일반행렬식분포(general determinantal distribution)으로서 $K = 1$, $\Lambda = (n-p)^{-1}$일 경우 아래와 같이 쓸 수 있다.

$$g(L) = \frac{C_{p,V}\pi^{-pK/2}|\Sigma|^{V/2}|\Lambda|^{p/2}}{C_{p,n}|\Sigma + (L-\Delta)\Lambda(L-\Delta)^T|^{n/2}}. \tag{4.14}$$

단, $v = n - K$이고, $C^{-1}{}_{p,v} = \pi^{p(p-1)/4}\prod_{j=1}^{p}\Gamma[1/2\cdot(v+1-j)]$이다.

M_a의 PSR 의사 베이즈 우도 L_a는 일반행렬식분포의 곱으로 나타낼 수 있다. 즉,

$$\begin{aligned} L_a &= \prod_{i=1}^{2}\prod_{j=1}^{n_i} g(Y_{ij},\ \Pi_i,\ M_a) \\ &= \prod_{t=1}^{m}\prod_{i=a(t)}^{b(t)}\prod_{j=1}^{n_i} D\{X\hat{\tau}_i F_{(j)},\ I,\ S^*(k_t),\ N(t)+k_t-1\}. \end{aligned} \tag{4.15}$$

여기서,

$$S(k_t)^* = \sum_{u=a(t)u\neq i}^{b(t)} \left\{(n_u - 1)S_u + \sum_{s}^{(j)}(y_{is} - X\tau_i(j)A_{(j)})(y_{is} - X\tau_i(j)A_{(j)})^T\right\}$$

$Y(ij)$: i번째 자료행렬 Y_i에 대해j번째 관측벡터 $y_{(ij)}$가 삭제된 자료이고,

t : 등분산 그룹의 수

$i = a(t),\ \cdots,\ b(t)$: 동일 그룹내 등분산의 수

j : 해당 모집단의 자료 수

$$n(t) = \sum_{i=a(t)}^{b(t)} n_i, \quad n = \sum_{i=1}^{2} n_i$$

이다.

둘째, $\Sigma_1 \neq \Sigma_2$의 경우 사전확률분포를 $h(\Sigma_i^{-1}) \propto |\Sigma|^{(p+1)/2}$이라 하면, M_b모형에서 다음과 같이 사후확률분포를 유도할 수 있다.

$$g(\Sigma_i^{-1}|Y_i) \propto |\Sigma_i|^{(n_i-p-1)/2}\exp\{-1/2tr\Sigma_i^{-1}(Y_i - X\hat{\tau}_i A_i)(Y_i - X\hat{\tau}_i A_i)\}. \tag{4.16}$$

이 사후확률분포로부터 아래 식을 얻을 수 있다.

$$E(\Sigma_i|Y_i) = (n_i - p - 1)^{-1}(Y_i - X\hat{\tau}_i A_i)(Y_i - X\hat{\tau}_i A_i)^T \tag{4.17}$$

이 때, 미래 관측치 y에 대한 예측분포는 다음과 같다.

$$G(y|\Sigma_i,\ Y_i,\ \Pi_i) = D\Big(X\tau_i F_i,\ I,\ (Y_i - X\hat{\tau}_i A_i)(Y_i - X\hat{\tau}_i A_i)^T,\ n_i + K\Big) \tag{4.18}$$

이 때, 모형 M_b의 PSR 의사 베이즈 우도 L_b도 역시 일반행렬식분포의 곱으로 나타낼 수 있다. 즉,

$$\begin{aligned} L_b &= \prod_{i=1}^{2}\prod_{j=1}^{n_i} g(y_{ij}|Y(ij),\ \Pi_i,\ M_b) \\ &= \prod_{t=1}^{m}\prod_{i=a(t)}^{b(t)}\prod_{j=1}^{n_i} D\Big\{X\hat{\tau}_{i(j)}F_{(j)},\ I, \\ &\qquad (Y_{i(j)} - X\hat{\tau}_{i(j)}A_{(j)})(Y_{i(j)} - X\hat{\tau}_{i(j)}A_{(j)})^T,\ n(t) + k_t - 1\Big\} \end{aligned} \tag{4.19}$$

식 (4.15)와 (4.19)로부터 성장곡선모형의 PSR 의사 베이즈 우도를 계산해 낼 수 있으며, Geisses(1979)등의 PSR 의사 베이즈 우도 기준법에 따라 $L_* = Max\{L_a,\ L_b\}$에 대응하는 의사 우도를 가진 모형 M^*를 선택할 수 있다.

5. 예 제

성장곡선모형에서 이상점의 존재가 등분산성의 검정에 좋지 않은 영향을 주는 경우, 이를 탐지하여 삭제한 후 등분산성검정을 행하는 것이 옳은 결과를 가져다 줄 수 있는 가를 알아 보기 위해 Potthoff등(1964)이 사용한 자료를 이용하여 예를 들어보았다.

이 자료는 Potthoff등이 처음 사용하였고, 후에 Lee(1991)등이 분석에 이용하였는데, 아동 들의 나이가 8, 10, 12, 14세로 변화함에 따라 뇌하수체 중심에서 익돌상악부위의 균열부 (pterygomaxillary fissure) 사이의 측정된 거리(단위 : mm)의 변화를 나타낸 자료이다.

(표-1) 아동들의 치아 측정치에 관한 자료(Potthoff등(1964))

	8세	10세	12세	14세
표본1(소녀)				
1	21	20	21.5	23
2	21	21.5	24	25.5
3	20.5	24	24.5	25.5
4	23.5	24.5	25	26.5
5	21.5	23	22.5	23.5
6	20	21	21	22.5
7	21.5	22.5	23	25
8	23	23	23.5	24
9	20	21	22	21.5
10	16.5	19	19	19.5
11	24.5	25	28	28
표본2(소년)				
12	26	25	29	31
13	21.5	22.5	23	26.5
14	23	22.5	24	27.5
15	25.5	27.5	26.5	27
16	20	23.5	22.5	26
17	24.5	25.5	27	28.5
18	22	22	24.5	26.5
19	24	21.5	24.5	25.5
20	23	20.5	31	26
21	27.5	28	31	31.5
22	23	23	23.5	25
23	21.5	23.5	24	28
24	17	24.5	26	29.5
25	22.5	25.5	25.5	26
26	23	24.5	26	30
27	22	21.5	23.5	25

측정값들이 동일한 시간간격으로부터 얻어졌으므로, 계획행렬 X와 A는 다음과 같다.

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 8 & 10 & 12 & 14 \end{pmatrix}^T, \quad A = \begin{pmatrix} I_{11} & 0 \\ 0 & I_{16} \end{pmatrix}$$

여기서, I_s는 1을 요소로 갖는 s-variant vector이다.

5.1 이상점탐지

식 (2.20)을 이용하여 상위 5개의 T_i 값을 계산하면 아래와 같다.

(표5-2) 치아자료에 대한 진단통계량

개체번호	T_i
24	1.9197
15	1.4433
21	1.2961
10	1.2738
20	1.2297

$\alpha = 0.01$라 하고 식 (3.3)을 사용하면

$$1 + \frac{m}{(n-r-p)C_\alpha} = 1.5505$$

이므로 24번째 개체만이 $T_{24} = 1.9197 > 1.5505$가 되어, 유의수준 1%하에서 이상점이라 할 수 있다.

5.2 등분산성 검정

SAS/IML을 이용하여 성장곡선모형에 대한 등분산성검정을 실시하였다. 표본1과 표본2에 대해 PSR 의사 베이즈 우도의 대수값($\log L^*$)을 다음의 모형 하에서 계산하였다.

$$M_a : \Sigma_1 = \Sigma_2$$
$$M_b : \Sigma_1 \neq \Sigma_2$$

이상점으로 탐지된 24번째 자료를 포함한 모든 자료를 사용하여 (식 4.15)와 (식 4.19)에 따라 PSR 의사 베이즈 우도값을 구하여 대수를 취하여 보면 $\log L_a = -17.94947$이 고, $\log L_b = -18.09958$을 얻을 수 있다. 이 결과 PSR 의사 베이즈 우도기준은 "$\Sigma_1 = \Sigma_2$"인 모형 M_a를 실제모형으로 선택하였음을 알 수 있다. 그러나, 24번째 자료를 제외한 후 계산한 결과 $\log L_a = -17.36215$이고, $\log L_b = -17.06154$가 되어 "$\Sigma_1 \neq \Sigma_2$"인 모형 M_b를 선택하여 위의 결과와는 달리 등분산성을 만족하지 않음을 알 수 있다. T_i값이 커 이상점으로 의심받는 15번째 자료를 24번째 자료와 함께 삭제한 후 계산한 결과 $\log L_a = -17.05683$이고, $\log L_b = -16.5276$이 되어 "$\Sigma_1 \neq \Sigma_2$"인 모형 M_b를 선택하고 있음을 알 수 있다. (표 5-3)에서는 (표 5-2)에서 제시한 T_i값이 큰 상위 5개 자료들을 순차적으로 제거한 후 계산한 $\log L_a$와 $\log L_b$값을 계산한 결과들을 나타내었다.

(표 5-3) T_i 값의 크기순으로 개체를 제거했을 경우의 PSR 의사 베이즈 우도의 대수값

제거한 자료의 개체번호	$\log L_a$	$\log L_b$
없음	-17.94947	-18.09958
24	-17.36215	-17.06154
24 15	-17.05683	-16.52760
24 15 21	-16.82780	-16.79422
24 15 21 10	-17.01896	-16.63337
24 15 21 10 20	-16.50164	-16.02664

(표 5-3)에서 보는 바와 같이 이상점으로 판단된 자료를 포함했을 경우를 제외한 모든 경우에 있어 $\log L_a$에 비해 $\log L_b$의 값이 크므로 "$\Sigma_1 \neq \Sigma_2$"인 모형 M_b를 선택하고 있음을 알 수 있다.

6. 결 론

본 논문에서는 평균이동모형을 사용하여 성장곡선모형에서 이상점을 탐지하는 방법을 제안하였다. 그리고 제안된 방법으로 이상점을 탐지하였을 경우, 이 점의 포함여부에 따라 그룹간의 등분산성 검정에 긍정적인 영향을 미치는가에 대해 PSR 의사 베이즈 우도 기준법을 적용하여 알아보았다.

그 결과 성장곡선모형에서 평균이동모형을 적용하여 탐지된 이상점을 제거한 후 등분산성 검정을 실시하는 것이 타당하다는 것을 알았다.

그러나, 성장곡선모형에서 다중 이상점을 효과적으로 탐지해 내는 문제와 이상점들 사이의 상호관계, 등분산성 검정에서 PSR 의사 베이즈 우도기준법 이외의 더 효과적인 방법이 존재하는 가의 여부는 다음의 연구과제로 남기고자 한다.

참 고 문 헌

1. Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression,* Wiley, New York.

2. Cook, R. D. and Weisberg, S. (1982). *Detection of influential observations in linear regression,* Chapman and hall, New York.

3. Hawkins, D. M. and McLachlan, G. J. (1997). High-Breakdown linear discriminant analysis, *Journal of American Statistical Association,* Vol.92, No.437, 136-143.

4. Geisser, S. (1970). Bayesian analysis of growth curves, *Sankhya Series.* A. 32,53-64.

5. Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection, *Journal of American Statistical Association,* Vol.74, 153-160.

6. Lee, L. C. and Geisser, S. (1975). Applications of growth curve prediction, *Sankya,* 37, 239-256.

7. Lee, J. C. (1991). Tests and model selection for the general growth curve model, *Biometrics,* 47, 147-159.

8. Liski, E. P. (1991). Detecting influential measurements in a growth curve model, *Biometrics,* 47, 659-668.

9. Pan, J. X. and Fang, K. T. (1995). Multiple outlier delection in growth curve model with unstructured covariance matrix, *Annals of the institute of Statistical Mathematics,* 47, 137-153.

10. Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems, *Biometrika,* 51,313-326.

11. Shim, K. B. (1993). A study for the equality test of covariance matrices by PSR quasi-Bayes criterion in small sample, Dongguk University.

12. von Rosen, D. (1989). Maximum likelihood estimates in multivariate linear model, *Journal of Multivariate Analysis,* 31,187-200.

13. von Rosen, D. (1990). Moments for a multivariate linear normal models with application to growth curve model, *Journal of Multivariate Analysis,* 35,243-259.

Outlier Detection in Growth Curve Model Using Mean-Shift Model [3]

Kyu-Bark, Shim [4]

Abstract

For the growth curve model with arbitrary covariance structure, known as unstructured covariance matrix, the problems of detecting outliers are discussed in this paper. In order to detect outliers in the growth curve model, the likelihood ratio testing statistics in mean shift model is established and its distribution is derived. After we detected outliers in growth curve model, we test homo and/or hetero-geneous covariance matrices using PSR Quasi-Bayes Criterion. For illustration, one numerical example is discussed, which compares between before and after outlier deleting.

Key Words and Phrases: Growth Curve Model, Outlier, Mean-Shift Model, Predictive Sample Reuse(PSR)

[3]This paper was supported by Dongguk University Research Fund 1999.

[4]Assistant Professor, Department of Statistics and information Science, Dongguk University, Kyongju, 780-714, Korea