

Outlier Detection in Growth Curve Model¹⁾

Kyu-Bark Shim²⁾

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 test statistics using U -distribution is established. After detecting 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.

Keywords : 성장곡선모형, Lawley-Hotelling Trace, U -분포

1. 머리말

성장곡선모형(growth curve model)은 분산분석에서 많이 이용되는 일반화된 다변량 모형이다. 이 모형은 동물과 식물의 성장에 관한 모형으로 특히 유용하다. 이 모형은 Potthoff 등(1964)에 의해 처음 제안된 후 많은 학자들에 의해 연구되어 왔다. 성장곡선모형은

$$Y_{p \times n} = X_{p \times m} \tau_{m \times r} A_{r \times n} + \epsilon_{p \times n} \quad (1-1)$$

과 같은 형태를 지녔는데, 여기서 X 와 A 는 계수가 각각 $m < p$, $r < n$ 인 이미 알려진 계획행렬(design matrix)이고, 모수 τ 는 모르는 수이다. 또한, 오차행렬 ϵ 의 열들은 평균벡터 0 이고, 미지의 합동공분산행렬 Σ 를 가진 독립된 p 차원 정규분포를 따른다.

일반적으로 성장곡선은 시간의 변화에 따른 다항식으로 나타나기 때문에 계획행렬

1) 이 논문은 2003학년도 동국대학교의 학술연구비 지원을 받아 연구되었음.

2) 경북 경주시 석장동 707번지 동국대학교 정보통계학과 부교수
E-mail : gpshim@dongguk.ac.kr

X 는 시간의 변화에 대응하여 얻어진 값들의 행렬이다. 또한 행렬 A 는 $p \times 1$ 차원 단위벡터(unit vector) $E_i, i = 1, 2, \dots, r$ 와 영벡터(null vector)들의 조합으로 이루어진다.

$$A_{r \times n}^T = I_r \otimes E_{pn} \quad (1-2)$$

여기서, E_{pn} 은 모든 단위 원소들의 $p \times n$ 행렬이고, \otimes 은 Kronecker 곱이다.

정규성의 가정 하에서 Rosen(1989)은 모수 τ 와 Σ 의 최우추정값(maximum likelihood estimate : MLE)들을 각각 아래와 같이 계산하였다.

$$\hat{\tau} = (X^T S^{-1} X)^{-1} X^T S^{-1} Y A^T (A A^T)^{-1} \quad (1-3)$$

$$\hat{\Sigma} = n^{-1} (Y - X \hat{\tau} A) (Y - X \hat{\tau} A)^T \quad (1-4)$$

여기서,

$$S = Y (I - A^T (A A^T)^{-1} A) Y^T \quad (1-5)$$

일반적인 선형모형 하에서 통계적 진단(statistical diagnostic) 문제에 대해서는 많은 학자들에 의해 회귀분석분야에서 이상점(outlier) 및 영향관측값(influential observation)을 탐지하는 주제로 연구되어 왔다. 그러나, 성장곡선모형에서 다중 이상점들이나 영향관측값들을 탐지하는 문제는 선형회귀모형에서의 문제에 비해 매우 복잡하여 많이 이루어지지 않고 있는 실정이다. 최근 Liski(1992)는 임의의 행렬 G 가 양정치행렬이고, $\sigma^2 > 0$ 가 미지의 스칼라일 때, 공분산행렬 $\Sigma = \sigma^2 G$ 를 가진 성장곡선모형에서 이상점들과 이상점들을 탐지하는 방법을 제안하였다.

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 이상점의 존재여부를 U -분포를 이용하여 판단하여 보고, 이상점의 제거가 등분산성 검정에 긍정적인 영향을 미치는지 알아보았다.

2. 이상점 진단

$I = \{i_1, i_2, \dots, i_k\}$ ($n > p + k$)를 제거해야 할 k 개의 개체들에 대한 첨자를 포함한 집합이라 하자. 보편성의 상실이 없다면, 첨자집합들은 $I = \{n - k + 1, n - k + 2, \dots, n\}$ 로 바꾸어 쓸 수 있으며, Y 는 $Y = (Y_{(I)}, Y_I)$ 로 분할할 수 있다. 여기서, $Y_{(I)}$ 는 식 (1-1)에서 Y_I 를 제거한 후의 성장곡선모형이다. 마찬가지로 행렬 A 와 ε 는 $A = (A_{(I)}, A_I)$ 및 $\varepsilon = (\varepsilon_{(I)}, \varepsilon_I)$ 로 각각 분할 할 수 있

다. 따라서, Y_I 를 제거한 후의 성장곡선모형 식 (1-1)은 아래와 같이 쓸 수 있다.

$$Y_{(I)} = X\tau A_{(I)} + \varepsilon_{(I)} \quad (2-1)$$

이 때, $\varepsilon_{(I)} \sim N_{p, n-k}(0, \Sigma \otimes I_{n-k})$ 이다. 같은 방법으로 식(2-1)에 대한 τ 와 Σ 의 MLE는 각각

$$\hat{\tau}_{(I)} = (X^T S_{(I)}^{-1} X)^{-1} X^T S_{(I)}^{-1} Y_{(I)} A_{(I)}^T (A_{(I)} A_{(I)}^T)^{-1} \quad (2-2)$$

$$\hat{\Sigma}_{(I)} = (n-k)^{-1} (Y_{(I)} - X\hat{\tau}_{(I)} A_{(I)}) (Y_{(I)} - X\hat{\tau}_{(I)} A_{(I)})^T \quad (2-3)$$

이다. 여기서,

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

Chatterjee 등(1988)은 식 (1-3)과 (2-2)를 이용하여 $\hat{\tau}$ 와 $\hat{\tau}_{(I)}$ 사이의 관계를 아래와 같이 도출한 바 있다.

$$\hat{\tau}_{(I)} = \hat{\tau} - (X^T S^{-1} X)^{-1} X^T S^{-1} e_I V_I^{-1} K_I^{-1} (A A^T)^{-1} \quad (2-4)$$

여기서 $V_I = I_k - H_I - e_I^T S^{-1} e_I + e_I^T S^{-1} X (X^T S^{-1} X)^{-1} X^T S^{-1} e_I$

$$H_I = A_I^T (A A^T)^{-1} A_I$$

$$K_I = A_I - A Y^T S^{-1} e_I + A Y^T S^{-1} X (X^T S^{-1} X)^{-1} X^T S^{-1} e_I$$

이고, $e = (e_{(I)}, e_I)$ 는 A 상에 회귀된 Y 의 잔차(residual)이다.

Cook(1982)에 의해 제안된 영향력 측정 방법은 선형회귀분석에서는 가장 많이 사용되나 성장곡선 모형에서는 사용할 수 없다. Liski (1992)는 성장곡선모형에서 영향력 상태에 대한 측정값을 도출해 내기 위해 다음과 같이 χ^2 통계량을 사용하였다.

$$\chi^2 = tr[(\hat{\tau} - \tau_*)^T A^T A (\hat{\tau} - \tau_*) X^T \Sigma^{-1} X] \quad (2-5)$$

여기서, τ_* 는 주어진 값이다. 식(2-5)에서 Σ 의 추정값을 사용하기 위해 $\Sigma^{-1} = \nu_E S^{-1}$ 를 사용하면, 다음과 같이 Lawley-Hotelling Trace 통계량을 구할 수 있다.

$$T_0^2 = \nu_E tr(S_H S_E^{-1}) = \nu_E U^{(s)} \quad (2-6)$$

여기서, $S_E = (X^T S^{-1} X)^T$

$$S_H = (\hat{\tau} - \tau_*)^T A^T A (\hat{\tau} - \tau_*)$$

이며, $s = \min(d, \nu_h)$ 인데, 이는 행렬 $S_H S_E^{-1}$ 의 0이 아닌 고유근의 수이다. d 는 다변량 정규분포의 차수이고, ν_h 와 ν_E 는 각각 S_H 와 S_E 의 자유도이다.

T_0^2 의 분포는 정확히 알 수 없으나, Seber(1984)가 정의한 $U^{(s)}$ 의 근사값을 이용하여 $100(1 - \alpha)\%$ 신뢰도에 대응하는 임계값을 추정할 수 있다. 즉, τ 에 대한 $100(1 - \alpha)\%$ 신뢰집합은 모든 τ 에 대해 다음을 만족하게 정의된다.

$$tr(S_H S_E^{-1}) \leq U_a^{(s)} \quad (2-7)$$

여기서, $U_a^{(s)}$ 는 $U^{(s)}$ 분포의 $100(1 - \alpha)\%$ 백분위수에 해당하는 지점의 점이다. Davis(1980)는 다양한 d 값에 대해 $\alpha = 0.05$ 및 0.01 부근에 해당하는 값들을 계산한 바 있다. 특히, McKeon(1974)은 $U^{(s)}$ 분포는 $F \sim F_{a, b}$ 일 때 cF 에 근사한다는 것을 보인 바 있다. 여기서,

$$\begin{aligned} a &= d\nu_H \\ b &= 4 + (a + 2) / (B - 1) \\ c &= a(b - 2) / b(\nu_E - d - 1) \\ B &= (\nu_E + \nu_H - d - 1) / (\nu_E - d - 3)(\nu_E - d) \end{aligned} \quad (2-8)$$

이다.

I 지점의 자료가 삭제되었을 경우의 영향력을 측정하기 위해 식 (2-5) 및 (2-6)에서 정의한 S_H 에서 τ_* 대신 $\tau_{(I)}$ 를 사용할 수 있다. 즉, 측정치 Y_I 를 삭제한 효과를 알기 위해 Y 와 $Y_{(I)}$ 의 대응되는 행들 사이의 Mahalanobis 거리의 제곱합을 이용하여 보자.

$$\begin{aligned} D_{(I)} &= tr(S_H S_E^{-1}) \\ &= tr[A(\hat{\tau} - \hat{\tau}_{(I)})X^T S^{-1} X(\hat{\tau} - \hat{\tau}_{(I)})^T A^T] \end{aligned} \quad (2-9)$$

식 (2-7)과 식 (2-9)의 관계에서 $(1 - \alpha) \times 100\%$ 신뢰도에 대응하는 임계값을 추정할 수 있으며, 식 (2-7)과 관련하여 아래 영역에 포함되는 점 I 에 대응하는 관측치는 영향력이 크며, 이를 이상점이라 할 수 있다.

$$D_{(I)} > cF(\alpha; a, b) \quad (2-10)$$

3. 성장곡선모형에서 등분산성 검정

Shim(1993)은 2개의 성장곡선모형의 경우 예측표본재이용(Predictive Sample Reuse : 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.

$$H_0: \Sigma_1 = \Sigma_2 \quad \text{대} \quad H_a: \Sigma_1 \neq \Sigma_2 \quad (3-1)$$

위의 가설 하에서 모집단 모형을 각각

$$M_a: \Sigma_1 = \Sigma_2 \quad \text{대} \quad M_b: \Sigma_1 \neq \Sigma_2 \quad (3-2)$$

라 하고, τ_i 의 값은 미지이나 추정 가능한 값이라 하자.

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

$$g(\Sigma^{-1} | Y) \propto |\Sigma|^{(n-p-1)/2} \exp\{-1/2 \text{tr} \Sigma^{-1} (Y - X\hat{\tau}A)(Y - X\hat{\tau}A)^T\}. \quad (3-3)$$

이 사후확률분포로부터 아래 식을 얻을 수 있다.

$$E(\Sigma | \tau, Y) = (n-p-1)^{-1} (Y - X\hat{\tau}A)(Y - X\hat{\tau}A)^T \quad (3-4)$$

이 때, 미래 관측치 y 에 대한 예측분포는 다음과 같다.

$$G(y | \tau, Y, \Pi_i) = D(X\tau F_i, I, (Y - X\hat{\tau}A)(Y - X\hat{\tau}A)^T, n + K) \quad (3-5)$$

여기서, 임의의 행렬 $L_{p \times K}$ 에 대하여 확률밀도함수가 식 (3-5)와 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.

$$g(L) = \frac{C_{p,v} \pi^{-pK/2} |\Sigma|^{v/2} |\Lambda|^{p/2}}{C_{p,n} |\Sigma + (L - \Delta) \Lambda (L - \Delta)^T|^{n/2}}. \quad (3-6)$$

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

M_a 의 PSR 의사 베이즈 우도 L_a 는 일반행렬식분포의 곱으로 나타낼 수 있다. 즉,

$$\begin{aligned}
L_a &= \prod_{i=1}^2 \prod_{j=1}^{n_i} g(Y_{ij}, \Pi_i, M_a) \\
&= \prod_{t=1}^m \prod_{i=a(t)}^{b(t)} \prod_{j=1}^{n_i} D\{X\hat{\tau}_i F_{(j)}, I, S^*(k_t), N(t) + k_t - 1\}. \quad (3-7)
\end{aligned}$$

여기서,

$$S(k_t)^* =$$

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

$Y(ij)$: i 번째 자료행렬 Y_i 에 대해 j 번째 관측벡터 $y_{(ij)}$ 가 삭제된 자료이고,

t : 등분산 그룹의 수

$i = a(t), \dots, b(t)$: 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.

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

$$g(\Sigma_i^{-1} | Y_i) \propto |\Sigma_i|^{(n_i - p - 1)/2} \exp\left\{-1/2 \operatorname{tr} \Sigma_i^{-1} (Y_i - X\hat{\tau}_i A_i)(Y_i - X\hat{\tau}_i A_i)^T\right\}. \quad (3-8)$$

이 사후확률분포로부터 아래 식을 얻을 수 있다.

$$E(\Sigma_i | Y_i) = (n_i - p - 1)^{-1} (Y_i - X\hat{\tau}_i A_i)(Y_i - X\hat{\tau}_i A_i)^T \quad (3-9)$$

이 때, 미래 관측치 y 에 대한 예측분포는 다음과 같다.

$$G(y | \Sigma_i, Y_i, \Pi_i) = D(X\tau_i F_i, I, (Y_i - X\hat{\tau}_i A_i)(Y_i - X\hat{\tau}_i A_i)^T, n_i + K) \quad (3-10)$$

이 때, 모형 M_b 의 PSR 의사 베이즈 우도 L_b 도 역시 일반행렬식분포의 곱으로 나타낼 수 있다. 즉,

$$L_b = \prod_{i=1}^2 \prod_{j=1}^{n_i} g(y_{ij} | Y(ij), \Pi_i, M_b)$$

$$= \prod_{t=1}^m \prod_{i=a(t)}^{b(t)} \prod_{j=1}^{n_i} D\{ X \hat{\tau}_{i(j)} F_{(j)}, I(Y_{i(j)} - X \hat{\tau}_{i(j)} A_{(j)}) \cdot (Y_{i(j)} - X \hat{\tau}_{i(j)} A_{(j)})^T, n(t) + k_t - 1 \} \quad (3-11)$$

식 (3-7)과 (3-11)로부터 성장곡선모형의 PSR 의사 베이지 우도를 계산해 낼 수 있으며, Geissess(1979)등의 PSR 의사 베이지 우도 기준법에 따라 $L_* = \text{Max}\{L_a, L_b\}$ 에 대응하는 의사 우도를 가진 모형 M^* 를 선택할 수 있다.

4. 예 제

성장곡선모형에서 이상점의 존재가 등분산성의 검정에 좋지 않은 영향을 주는 경우, 이를 탐지하여 삭제한 후 등분산성검정을 행하는 것이 옳은 결과를 가져다 줄 수 있는가를 알아보기 위해 Weissfeld 등(1992)이 사용한 성장곡선 자료를 이용하여 예를 들어보았다.

표 <4-1>의 자료는 임상연구 결과 얻어진 것인데, 집중적인 감시를 요하는 환자들에게 기관내의 흡입술 가운데 표준 흡입술 및 새로운 흡입술을 실시한 결과이다.

이 방법에 따라 산소침윤(oxygen saturation)의 결과를 흡입하기 전(시점1), 첫 번째 흡입후(시점2), 두 번째 흡입 후(시점3), 세 번째 흡입 후(시점4), 마지막 흡입으로부터 5분 후 (시점5)등 5개 시점에서 측정하였다. 실험자는 비록 동일한 환자들을 대상으로 측정하였으나, 실험에 사용된 두 가지 흡입술이 서로 다른 방법이므로 측정 자료들 사이의 등분산성은 가정하지 않는 것이 옳다고 믿었다.

4-1. 이상점탐지

값들의 정확한 측정시점을 알 수 없으므로 동일한 시간 간격이라 간주하면, 계획행렬 X 와 A 는 다음과 같이 쓸 수 있다.

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 5 & 10 & 15 & 20 \end{pmatrix}^T, \quad A^T = I_2 \otimes E_{5, 25}$$

U -분포를 F 분포에 근사 시키기 위해 식 (2-8)을 이용하자. 여기서, $d = 2$, $\nu_H = 2$, $d_E = 48$ 이므로, $a = 4$, $b = 10$, $c = 0.071$ 를 얻을 수 있다. 식 (2-9)에 대해 SAS/IML을 이용하여 상위 5개의 $D_{(I)}$ 값을 계산하면 <표 4-2>와 같다.

$\alpha = 0.05$ 라 하고 F 분포에 근사시키면, $cF(0.05; 2, 10) = 0.2911$ 이므로 표준방법의 경우 25번째 개체와 새로운 방법의 경우 5번째, 24번째 및 5번째 개체가 각각 0.2911보다 큰 값을 가지게 되어, 유의수준 5%하에서 이상점이라 할 수 있다.

4-2. 등분산성 검정

‘표준 흡입술’ 및 ‘새로운 흡입술’에 대해 PSR 의사 베이지 우도의 대수값($\log L^*$)을 다음의 모형 하에서 계산하였다.

$$M_a: \Sigma_1 = \Sigma_2 \quad \text{대} \quad M_b: \Sigma_1 \neq \Sigma_2$$

이상점으로 탐지된 자료를 포함한 모든 자료를 사용하여 식 (3-7)과 (3-11)에 따라 검정통계량을 구하여 대수를 취하여 보면 $\log L_a = -28.0738$ 이고, $\log L_b = -26.3171$ 을 얻을 수 있다. 이 결과 “ $\Sigma_1 = \Sigma_2$ ”인 모형 M_a 를 실제모형으로 선택하였음을 알 수 있다. 그러나, 이상점으로 탐지된 표준 방법의 25번째 자료와 새로운 방법의 24번째 자료 및 5번째 자료를 순차적으로 제외한 후 계산한 결과는 각각 -24.9961 이 되어 “ $\Sigma_1 \neq \Sigma_2$ ”인 모형 M_b 를 선택하므로 위의 결과와는 달리 등분산성을 만족하지 않는 것이 옳음을 알 수 있다. <표 3-3>에서는 <표 3-2>에서 제시한 이상점들을 순차적으로 제거한 후 계산한 $\log L_a$ 와 $\log L_b$ 값을 계산한 결과들을 나타내었다.

<표 4-3>에서 보는 바와 같이 이상점으로 판단된 자료를 제거한 모든 경우에 있어 $\log L_a$ 에 비해 $\log L_b$ 의 값이 크므로 “ $\Sigma_1 \neq \Sigma_2$ ”인 모형 M_b 를 선택하고 있음을 알 수 있다.

<표 4-2> 흡입술 자료에 대한 $D_{(I)}$ 값

표준 흡입술		새로운 흡입술	
개체번호	$D_{(I)}$	개체번호	$D_{(I)}$
25	0.5359	24	0.4564
1	0.2024	5	0.3738
24	0.1726	22	0.2235
5	0.1721	2	0.2094
22	0.1544	3	0.1830

<표 4-3> 이상점들을 순차적으로 제거했을 경우의 PSR 의사베이지우도의 대수값

제거한 자료의 개체번호		$\log L_a$	$\log L_b$
표준 흡입술	새로운 흡입술		
없음	없음	-28.0738	-26.3171
25	24	-22.9376	-24.9236
25	24 5	-23.1861	-24.9961

<표 4-1> 환자들에게 두 가지 흡입술을 실시하여 얻은 자료

환자 번호	시점1		시점2		시점3		시점4		시점5	
	표준 흡입술	새로운 흡입술	표준 흡입술	새로운 흡입술	표준 흡입술	새로운 흡입술	표준 흡입술	새로운 흡입술	표준 흡입술	새로운 흡입술
1	95	94	96	95	94	95	97	95	95	94
2	94	96	94	96	92	95	93	95	95	94
3	94	92	93	94	92	93	91	94	93	92
4	96	98	98	98	97	99	98	98	95	97
5	94	94	93	90	94	93	95	93	95	95
6	97	98	99	97	100	98	99	94	99	98
7	97	95	90	95	93	96	91	93	97	97
8	94	93	95	96	95	96	95	94	95	95
9	95	97	96	99	95	100	94	100	94	99
10	96	100	96	100	96	100	96	100	96	100
11	98	99	96	99	99	98	97	98	98	99
12	94	97	89	98	88	98	89	98	93	96
13	93	93	93	94	94	95	94	95	95	94
14	97	100	100	100	97	99	98	99	96	100
15	96	96	97	95	99	95	100	96	97	97
16	92	92	93	90	94	91	93	96	93	93
17	100	96	99	95	98	97	99	97	99	96
18	97	98	96	100	96	100	96	100	96	99
19	94	98	97	97	98	96	98	96	96	100
20	94	96	97	96	97	96	95	96	95	97
21	96	97	96	95	94	95	94	92	96	96
22	97	99	98	95	97	93	96	90	98	99
23	97	100	98	100	98	100	99	100	98	100
24	98	90	100	91	100	92	100	91	96	89
25	90	94	90	93	90	94	96	94	90	94

$\log L_a = -22.9376$, $\log L_b = -24.9236$ 및 $\log L_a = -23.1861$, $\log L_b = -24.9961$

5. 결 론

본 논문에서는 U -분포를 이용하여 성장곡선모형에서 이상점을 탐지하는 방법을 소개하고 Weissfeld 등(1992)이 사용한 성장곡선 자료를 이 방법에 적용하여 보았다.

그리고 소개된 방법으로 이상점을 탐지하였을 경우, 이 점의 포함여부에 따라 그룹간의 등분산성 검정에 긍정적인 영향을 미치는 가에 대해 PSR 의사 베이즈 우도 기준법을 적용하여 알아보았다. 그 결과 성장곡선모형에서 탐지된 이상점을 제거한 후 등분산성검정을 실시하는 것이 타당하다는 것을 알았다.

그러나, 성장곡선모형에서 다중 이상점을 효과적으로 탐지해 내는 문제와 이상점들 사이의 상호관계, 더 효과적인 등분산성 검정방법에 대한 연구는 다음의 과제로 남겨 고자 한다.

참 고 문 헌

1. 심규박(2002). 성장곡선모형의 이상점 탐지에 U -분포의 적용, 한국데이터정보과학회 2002년 추계학술발표 논문집, 149-156.
2. Chatterjee, S. and Hadi, A.S.(1988). *Sensitivity Analysis in Linear Regression*, Wiley, New York.
3. Cook, R.D. and Weisberg, S.(1982). *Detection of influential observations in linear regression*, Chapman and hall, New York.
4. Hawkins, D.M. and McLachlan, G.J.(1997). High-Breakdown linear discriminant analysis, *Journal of American Statistical Association*, Vol.92, No.437, 136-143.
5. Davis, A.W(1980). Further tabulation of Hotelling's T_0^2 . *Comm. Stat. Simulation Computer. B*. 9. 321-336.
6. Geisser, S. and Eddy, W.F.(1979). A predictive approach to model selection, *Journal of American Statistical Association*, Vol.74, 153-160.
7. Liski, E.P.(1992). Detecting influential measurements in a growth curve model, *Biometrics*, 47, 659-668.
8. McKeon, J.J.(1974). F approximations to the distribution of Hotelling's T_0^2 , *Biometrika*, 61, 381-383.
9. 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.
10. von Rosen, D.(1989). Maximum likelihood estimates in multivariate linear model, *Journal of Multivariate Analysis*, 31, 187-200.
11. Seber, G.A.F.(1984). *Multivariate Observations*. John Wiley & Sons, Inc. New York.
12. Shim, K.B.(1993). *A study for the equality test of covariance matrices*

- by PSR quasi-Bayes criterion in small sample*, Dongguk University.
13. Weissfeld, L.A. and Kshirsagar, A.M.(1992). A modified growth curve model and its application to clinical studies, *Australian Journal of Statistics*, 34, 161-168.

[2003년 3월 접수, 2003년 5월 채택]