Journal of Korean
Data & Information Science Society
2002, Vol. 13, No.2 pp 185 193

Combination of Schwarz Information Criteria for Change-Point Analysis

Jong Tae Kim [1)]

Abstract

The purpose of this paper is to suggest a method for detecting the linear regression change-points or variance change-points in regression model by the combination of Schwarz information criteria. The advantage of the suggested method is to detect change-points more detailed when one compares the suggest method with Chen(1998)'s method.

Keywords : Linear regression fit; Change-point; Information criteria.

1. 서론

정보판별(Information Criterion)함수는 시계열분석, 이상점(outliers)의 분석, 회귀분석, 가설검정론 등 많은 통계 추론 분야에 사용되어져 왔다. 모수 θ와 모수의 개수 p와 최우도함수 $L(\hat{\theta})$에 대하여, Akaike(1973)에 의해 제시된 $-2\log L(\hat{\theta})+2p$인 정보판별함수(Akaike Information Criterion, AIC)와 $-2\log L(\hat{\theta})+p\log n$로 정의되는 Schwarz(1978)의 정보판별함수(Schwarz Information Criterion, SIC), 그리고 AIC 와 SIC를 기초로 변형된 많은 정보판별함수들은 경제학, 심리학, 공학 등, 결정과 판단이 필요한 많은 학문 분야에 유용한 도구로서 사용되어져 왔다. 특히 SIC는 일치추정량의 성질을 가지는 정보판별함수로서 변화점 추정의 도구로서 사용되었다.

Chen(1998)은 관찰된 데이터의 집합에서 선형회귀모형의 변화에 따른 변화점을 SIC를 이용하여 모형에 포함된 여러개의 변화점들 중 가장 영향력이 큰 한 개의 변화점에 관한 추정 방법을 제시하였다. Chen과 Gupta(1997)의 연구는 분산에 대한 변화점들을 검정 하는데 기초를 두지만 데이터에 포함된 변환점들을 찾기 위하여 이분법에 의한 방법으로 순차적인 방법으로 여러번의 반복 검정을 해야 하는 불편함이 있다.

실제로 어느 시점 x_0가 변화점이라고 가정할 때, 데이터에 포함되어 있는 모형함

---

1. Associate Professor, Dept of Statistics, Taegu University, 712-714, Korea.
E-mail: jtkim@daegu.ac.kr

수에 의하여 변화점이 발생되는 경우가 있고, 혹은 모형의 확률오차에 포함된 분산의 변화에 따른 변화점이 발생될 경우가 생긴다. 이들 두가지 중 어느 경우라도 관계없이 변화점 x_0 을 중심으로 이웃하고 있는 영역의 국소적인 부분에서의 분산의 변화가 필연적으로 발생하게 되는 경우를 실험을 통하여 볼 수 있다.

그러므로 본 연구의 아이디어는 모형의 확률오차에 포함된 분산의 변화에 대한 통계량과 선형회귀직선의 변화에 대한 통계량을 혼합하여 진보된 통계량을 제시하는데 있다. 제시된 통계량은 데이터의 변화점이 발생하는 위치들을 보다 더 정확하게 관찰할 수 있고, 추정된 변화점에 의해 관찰되는 선형회귀직선을 기준으로 데이터가 등분산 성질을 잘 나타내고 있다. 또한 회귀 모형의 변화에 대한 변화점이나 혹은 분산의 변화에 대한 변화점을 동시에 검정할 수 있는 효과를 가질 수 있다.

2. 정보 판별기준에 의한 변화점추정

다음의 선형(회귀)모형에 대하여 고려해 보자.

$$y_i = \mu_i + \varepsilon_i , \quad i = 1, \dots , n .$$

여기서 ε_i는 i번째 관측의 확률오차로서 평균이 0이고 분산 σ^2인 정규분포를 따른다고 가정하자. 따라서 이 모형의 i번째 반응값 y_i는 평균이 μ_{y_i}이고 분산이 σ^2인 정규모집단에서 독립적으로 나오는 값임을 의미한다. 만약 $\mu_{y_i} = \beta_0 + \beta_1 x_i$로 둔다면 다음의 단순선형 회귀모형이 성립할 것이다.

$$y_i = \mu_{y_i} + \varepsilon_i = \beta_0 + \beta_0 x_i + \varepsilon_i , \quad i = 1, \dots , n . \qquad (2.1)$$

본 논문에서 제시하는 변화점의 추정방법은 단순회귀에 국한하지 않고 중선형회귀모형의 확장에서도 쉽게 적용할 수 있다. 그러나 이해의 편의를 돕기위하여 단순 선형회귀모형에서 변화점 추정방법을 제시한다. 어떤 시점, 가령 관찰값들의 각각의 $k_1, k_2, \dots, k_{m-1}$번째 이후에 관찰값들이 절편이나 기울기의 변화에 따라 서로 다른 선형회귀모형들의 변화나 혹은 분산의 변화에 따라서 선형회귀모형의 변화를 심각하게 고려해야하는 속성을 가지는 데이터들의 집합을 흔히 접한다. 이에 대한 가설검정을 위해서 각각의 서로 다른 미지의 회귀모수 β_{0j}, β_{1j}, $j = 1, \dots, m$과 $k_0 = 1$, $k_m = n$에 대하여, $\mu^j_{y_i} = \beta_{0j} + \beta_{1j} x_i$, $i = k_{j-1}, \cdots, k_j$, $j = 1, \cdots m$으로 두고, 다음과 같은 귀무가설과 대립가설을 정의하자.

$$\text{귀무가설, } H_0 : y_i \sim N(\mu_{y_i}, \sigma^2) \quad i = 1, \cdots, n \qquad (2.2)$$

대립가설,

$$\begin{aligned} H_1 :\ & y^1_i \sim N(\mu^1_{y_i}, \sigma^2_1), \quad i = 1, \cdots, k_1, \\ & y^2_i \sim N(\mu^2_{y_i}, \sigma^2_2), \quad i = k_1 + 1, \cdots, k_2, \qquad (2.3) \\ & \cdots\cdots\cdots\cdots\cdots\cdots \\ & y^m_i \sim N(\mu^m_{y_i}, \sigma^2_m), \quad i = k_{m-1} + 1, \cdots, n . \end{aligned}$$

여기서 각각의 $k_1, k_2, \dots, k_{m-1}$은 상수의 값을 가지며 $1 < k_1 < k_2 < \cdots < k_{m-1} < n - 1$

의 범위를 가진다. 그리고 $k_j+1,\ j=1,\dots,m-1$은 선형회귀모형의 변환이 일어나는 변화점이 된다. 즉, 귀무가설 H_0는 데이터의 집합에서 어떠한 변화점을 가지지 않는 단일 선형회귀모형을 가지고 있음을 의미하고, 대립가설 H_1은 각각의 $k_j+1,\ j=1,\dots,m-1$의 변화점을 기준으로 m개의 회귀모형을 가짐을 뜻한다.

2.1 선형모형의 변화점추정

먼저 식 (2.3)의 다중 변화점에 대한 대립가설에서의 Schwarz의 정보판별함수 통계량의 성질을 조사하기 전에 한 개의 변화점이 존재하는 경우의 검정의 가설에 대한 SIC 통계량을 구해보자. Chen (1997)은 Schwarz의 정보판별함수를 이용하여, 회귀모형에서의 가장 영향력이 큰 하나의 변화점을 찾는 방법으로 귀무가설에서의 통계량을 다음과 같이 제시하였다.

$$\begin{aligned}\mathrm{SIC}_c(n) &= -\log L_0(\hat\beta_0,\ \hat\beta_1,\ \hat\sigma_c^{\,2}) + 3\log n \\ &= n\log\left(\hat\sigma_c^{\,2}\right) + n(1+\log 2\pi) + 3\log n.\end{aligned} \tag{2.4}$$

여기서 $\hat\beta_0$와 $\hat\beta_1$은 β_0와 β_1에 대한 최소제곱추정량이고, 이를 이용한 추정된 회귀직선 $\hat\mu_{y_i}=\hat\beta_0+\hat\beta_1 x_i$에 대하여, 분산 σ^2의 추정량은 $\hat\sigma_c^{\,2}=\frac{1}{n}(y-\hat\mu_{y_i})^2$이다.

다음은 대립가설 H_1에 대한 SIC 통계량을 조사한다. 데이터를 임의의 시점에서 이분화하여 생성된 미지의 회귀모수 β_{01}과 β_{11}, 그리고 β_{02}와 β_{12}와 σ^2에 대한 최우도 추정치는 다음과 같다.

$$\begin{aligned}\widehat{\beta_{01}} &= \overline{y_k} - \widehat{\beta_{11}}\,\overline{x_k}, \\ \widehat{\beta_{11}} &= \sum_{i=1}^{k}(x_i-\overline{x_k})(y_i-\overline{y_k}) \Big/ \sum_{i=1}^{k}(x_i-\overline{x_k})^2, \\ \widehat{\beta_{02}} &= \overline{y}_{n-k} - \widehat{\beta_{12}}\,\overline{x}_{n-k}, \\ \widehat{\beta_{12}} &= \sum_{i=1}^{k}(x_i-\overline{x}_{n-k}) \Big/ \sum_{i=1}^{k}(y_i-\overline{y}_{n-k}) \\ \widehat{\sigma_{c1}}^{\,2} &= \frac{1}{n}\left\{\sum_{i=1}^{k}(y_i-\widehat{\beta_{01}}-\widehat{\beta_{11}}x_i)^2 + \sum_{k=k+1}^{n}(y_i-\widehat{\beta_{02}}-\widehat{\beta_{12}}x_i)^2\right\}.\end{aligned}$$

여기서

$$\overline{x}_k = \frac{1}{k}\sum_{i=1}^{k}x_i,\quad \overline{y}_k=\frac{1}{k}\sum_{i=1}^{k}y_i \quad \text{이고,}$$

$$\overline{x}_{n-k} = \frac{1}{n-k}\sum_{i=k+1}^{n}x_i,\quad \overline{y}_{n-k}=\frac{1}{n-k}\sum_{i=k+1}^{n}y_i$$

이다. 그러므로, 대립가설 H_1에서 $k=2,\cdots,n-2$에 대한 Schwarz 정보판별함수는 최우도 함수의 추정값은 다음과 같이 구할 수 있다.

$$\begin{aligned} SIC_c(k) &= -2\log L_1(\widehat{\beta_{01}}, \widehat{\beta_{11}}, \widehat{\beta_{02}}, \widehat{\beta_{12}}, \widehat{\sigma_{c1}}^2) + 5\log n \\ &= n\log\{\widehat{\sigma_{c1}}^2\} + n(1+\log 2\pi) + 5\log n \end{aligned} \tag{2.5}$$

여기서 $L_1(\widehat{\beta_{01}}, \widehat{\beta_{11}}, \widehat{\beta_{02}}, \widehat{\beta_{12}}, \widehat{\sigma_{c1}}^2)$ **은 대립가설** H_1**하에서 최우도 함수이다. 선형회귀모형의 변화점 가설검정을 위한 판정기준은**

$SIC_c(n) \le \min\{SIC_c(k) : 2 \le k \le n-2\}$**이면 귀무가설** H_0**을 채택.** (2.6)

하고, 이와 반대로 $SIC_c(\hat{k}) = \min\{SIC_c(k) : 2 \le k \le n-2\} < SIC_c(n)$**이면 대립가설** H_0 **을 기각한다. 이 때 변화점 추정을 위한** $\hat{k}$ **는** $\{SIC_c(k) : 2 \le k \le n-2\}$ **최소로하는** k **이고 변화점은** $\hat{k}+1$**이 된다.**

2.2 분산의 변화점추정

독립변수에 따라 분산의 변화가 생길 때, SIC**에 의한 검정방법은 다음과 같다, 즉, 귀무가설** $H_0 : \sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2$**와 대립가설** $H_1 : \sigma_1^2 = \cdots = \sigma_{k_0}^2 \ne \sigma_{k_0+1}^2 = \cdots = \sigma_n^2$ **에 대하여, 귀무가설하에서의** SIC**의 정보판별함수 통계량은 다음과 같다.**

$$SIC_g(n) = n\log\widehat{\sigma_g}^2 + n(1+\log 2\pi) + 3\log n. \tag{2.7}$$

여기서 $\widehat{\sigma_g}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i)^2$**이다. 대립가설하에서의** SIC**의 통계량:**

$$SIC_g(k) = k\log\widehat{\sigma_{g1}}^2 + (n-k)\log\widehat{\sigma_{g2}}^2 + n(1+\log 2\pi) + 6\log n, \tag{2.8}$$

$\widehat{\sigma_{g1}}^2 = \frac{1}{k}\sum_{i=1}^{k}(y_i - \widehat{\beta_{01}} - \widehat{\beta_{11}}x_i)^2$, $\widehat{\sigma_{g2}}^2 = \frac{1}{(n-k)}\sum_{i=k+1}^{n}(y_i - \widehat{\beta_{02}} - \widehat{\beta_{12}}x_i)^2$**이다.**

위의 식 (2.8)**의** SIC_g**는** Chen**과** Gupta(1997)**가 제시한 추정량을 변형하여 제시한 추정량이다. 앞 절의 설명과 같이 어느 변화점을 중심으로 양쪽 데이터들의 각각에 대한 선형회귀모형에서의 평균제곱오차**(mean square error, MSE)**의 합을 추정하여** MSE**의 합들의 변화에 대한** SIC**의 값들을 관찰한 것이다. 이러한 측면에서** Chen**과** Gupta**의 추정량과 본 연구에서 제시하는 분산의 변화에 대한 추정량은 다르다. 제시된 식** (2.8)**의 분산의 변화에 대한 통계량을 이용하여 변화점을 추정하는 방법은 다음의 일반적인 절차를 따른다.**

$$SIC_g(\hat{k}) = \min\{SIC_g(k) : 2 \le k \le n-2\}. \tag{2.9}$$

여기서 $\hat{k}$ **는** $SIC_g(k)$**를 최소로하는 값이다. 검정에 대한 판단기준을 정하기 위하여, 신뢰계수** α**에 대하여 관계된 기각값을** $c_\alpha \ge 0$**이라고 하자. 만약**

$SIC_g(n) > \min\{SIC_g(k) : 2 \le k \le n-2\} + c_\alpha$**이라면 귀무가설** H_0 **을 기각** (2.10)

한다. 그리고 $\hat{k}+1$ **의 값을 변화점으로 간주한다. 그러므로 신뢰계수** α**에 관계된 기각값** c_α**에 대하여 다음의 확률이 성립된다.**

$$1-\alpha = P[SIC_g(n) < \min\{SIC_g(k),\ 2 \le k \le n-2\} + c_\alpha | H_0]. \tag{2.11}$$

식 (2.11)의 c_α의 값은 Gupta와 Chen(1997)의 연구결과와 비슷한 방법으로 구할 수 있다. 다음 절에서는 위의 두 통계량을 이용한 변화점의 추정 방법을 제시한다.

3. 제시된 변화점 추정

앞절에서 Schwarz의 정보판별함수를 이용한 분산의 변화에 따른 변화점 추정방법 SIC_g와 Chen에 의해 제시된 회귀직선의의 변화에 따른 변화점 추정 SIC_c을 관찰하였다.

앞서 설명한 바와 같이 실제로 어느 시점 x_0에서 모평균 변화에 의한 변화점이 발생되거나 혹은 회귀직선의 변화에 따른 변화점이 발생될 경우에, 그 변화점 x_0을 중심으로 이웃하고 있는 영역의 국소적인 부분에서의 분산의 변화가 필연적으로 발생하게 된다. 그러므로 분산의 변화에 대한 통계량 SIC_g와 회귀직선의 변화에 대한 통계량 SIC_c에 의한 변화점의 추정보다는 두 가지 방법을 혼합한 $TIC = (SIC_g + SIC_c)/2$에 의한 추정이 보다 더 세밀하게 변화점을 관찰할 수 있는 장점을 가진다. 귀무가설에서의 통계량 TIC 를 다음과 같이 정의하자.

$$\begin{aligned} TIC(n) &= (SIC_g(n) + SIC_c(n))/2 \\ &= \frac{n}{2}\left\{\log \widehat{\sigma_g}^2 + \log\left(\widehat{\sigma_c}^2\right)\right\} + n(1+\log 2\pi) + 3\log n. \end{aligned} \tag{3.1}$$

대립가설에서의 통계량:

$$\begin{aligned} TIC(k) &= (SIC_g(k) + SIC_c(k))/2 \\ &= \frac{k}{2}\log \widehat{\sigma_{g1}}^2 + \frac{(n-k)}{2}\log \widehat{\sigma_{g2}}^2 + \frac{n}{2}\log\left(\widehat{\sigma_{c1}}^2\right) \\ &\quad + n(1+\log 2\pi) + 5.5\log n. \end{aligned} \tag{3.2}$$

변화점 가설검정을 위한 판정기준은

$$TIC(n) \le \min\{TIC(k) : 3 \le k \le n-4\} \text{이면 귀무가설 } H_0\text{을 채택} \tag{3.3}$$

하고, 이와 반대로

$$TIC(\widehat{k}_1, \widehat{k}_2, \ldots,) = \text{local arg min}_{k_1, k_2, \ldots,}\{TIC(k) : 3 \le k \le n-4\} < TIC(n) \tag{3.4}$$

이면 H_0을 기각한다. 여기서 k의 범위는 $3 \le k \le n-4$이다. 이는 $TIC(k)$에 포함된 추정된 모수들의 수의 영향을 받았다. 식 (3.4)에서 local arg min $\{TIC(k) : 3 \le k \le n-4\}$의 표현은 $TIC(n)$의 값보다 작은 $TIC(k)$, $3 \le k \le n-4$, 값들이 여러 개의 국소적인 집단(cluster)영역에서 국소적인 최소값 $\widehat{k}_1, \widehat{k}_2, \ldots$을 갖는 것을 의미한다. 이들 국소적인 집단의 개수는 추정된 변화점의 개수로 추정되고, 국소적인 집단영역의 최소값들 $\widehat{k}_1, \widehat{k}_2, \ldots$에 대하여 다음 위치 $\widehat{k}_1+1, \widehat{k}_2+1, \ldots$가 변화점의 위치로

추정된다. 그러므로 신뢰계수 α에 관계된 기각값 c_α에 대하여 다음의 확률이 성립된다.

$$1-\alpha = P\left[\,TIC(n) < \text{local arg min}\{TIC(k),\ 3\le k\le n-4\} + c_\alpha \mid H_0\,\right]. \quad (3.5)$$

이러한 아이디어는 모집단내의 임의의 k 시점에서 모형의 변화가 발생한다면 k 지점의 국소적인 부분에서 분산의 변화가 발생하는 것과 동일한 성격을 가진다. 이것은 점프에 의한 변화이던지 분산이나 혹은 회귀직선의 변화에 의한 모형이든 상관없이 모두 통계량 TIC에 의해 관찰되어질 것이다. 그러므로 제시된 TIC는 분산의 변화점이나 회귀모형의 변화점이 둘 다 혼합된 모형에 대한 변화점 추론에 영향력을 가진다.

4. 실제 사례의 적용

식 (3.4)의 이해를 돕기위하여 그림 4.1은 Holbert(1982)의 논문에 있는 1967년 1월부터 1969년 11월까지, 반응변수로 보스톤 증권 회사(Boston Stock Exchange , BSE)의 매월 미화의 판매량과 설명변수로 뉴욕 어메리칸 증권회사(New York American Stock Exchange, NYAMSE)의 매월 미화의 판매량의 데이터이다.

그림 4.1

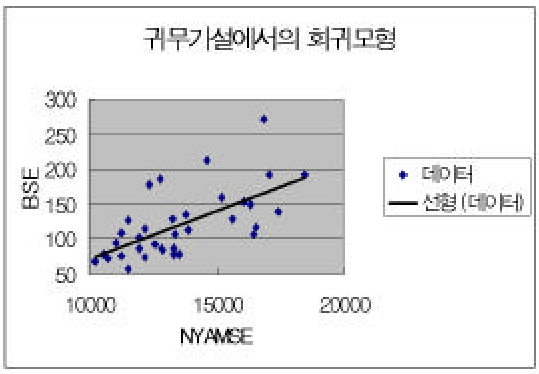

그림 4.2

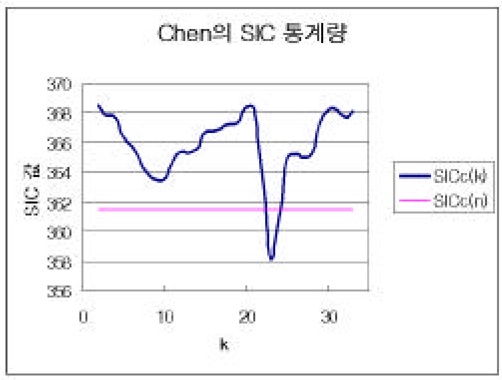

그림 4.1의 실선으로 된 부분은 귀무가설 H_0에서의 추정된 회귀직선을 나타낸 것이다. 그림 4.1에서 보듯이 데이터의 모형은 추정된 하나의 회귀직선을 기준으로 데이터는 등분산성의 성질을 만족하지 않음을 알 수 있다. 그림 4.2와 그림 4.3은 Holbert의 데이터를 가지고 분석한 Chen(1998)의 결과이다. Chen의 결과에 따르면 1968년 12월 (24번째 데이터) 이후의 NYAMSE에 대한 BSE의 선형모형의 변화가 일어나고 있음을 의미한다. 변화점(24번째 데이터)인 1968년 12월 이전과 이후의 데이터들을 분리하여 회귀모형을 분리하는 것이 보다 바람직함을 알 수 있다

그림 4.4은 Chen에 의해 두 개의 분리된 회귀모형을 나타낸 그림이다. 그림 4.4에서

모형(1-23)의 데이터의 집합은 추정된 회귀직선으로부터 NYAMSE의 값이 크질수록 BSE의 값의 산포가 점점 더 크게 나타나고 있다. 이것은 1에서 23째의 데이터들 사이에서 분산의 변화가 일어나고 있음을 나타낸다. 그러므로 분산의 변화에 대한 변화점의 추정이 필요하다.

그림 4.3

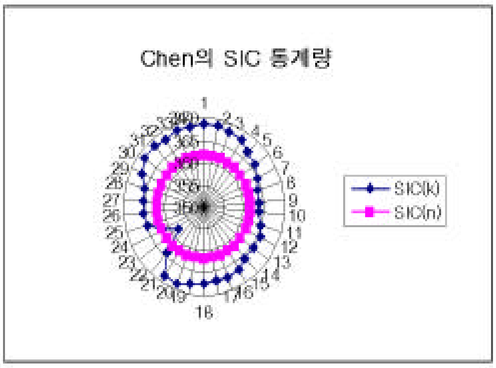

그림 4.4

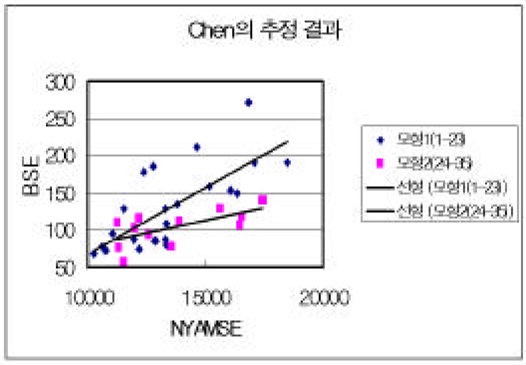

그림 4.5와 그림 4.6은 식 (3.1)과 식 (3.2)의 제시된 $TIC(n)$과 $TIC(k)$에 대한 그림이다. 식 (3.4)의 $TIC(\hat{k}) = \text{local arg min}\{TIC(k) : 2 \le k \le n-2\} < TIC(n)$의 기각영역에 따라서 $\hat{k}$ = 9, 14, 23 의 3가지 경우의 변화점 $\hat{k}+1$ = 10, 20, 24가 추정되어진다. 그러나 $\hat{k}=14$인 경우에, $TIC(k)$의 값과 $TIC(n)$값과의 거리의 차이가 작게 나타난다. 이에대한 해석은 실제로 가설검정의 기각역 C_α의 값의 결정을 요구한다.

그림 4.5

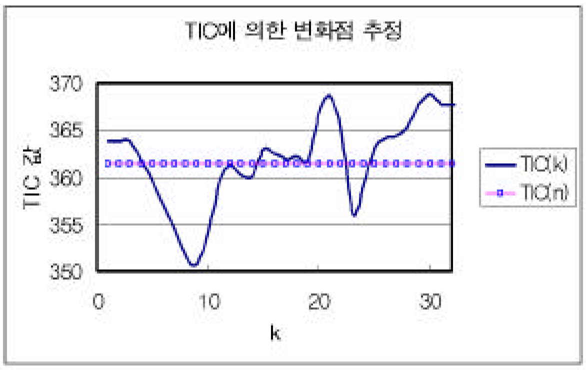

그림 4.6

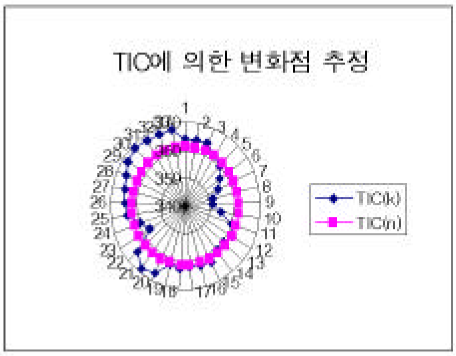

그림 4.7은 변화점이 3개인 경우 $\hat{k}+1 = 10, 20, 24$에 대하여 4개의 회귀적선들을 추정한 것이다. 그림 4.7에 나타난 선형회귀 모형들을 자세히 살펴보면 각각의 모형들의 데이터는 각 추정된 회귀직선을 기준으로 동일하게 분포되어 있음을 알 수 있

다. 또한 시간 k에 따른 각 추정된 회귀직선들은 서로 다른 모형을 가지고 변하고 있음을 알 수 있다.

그림 4.7

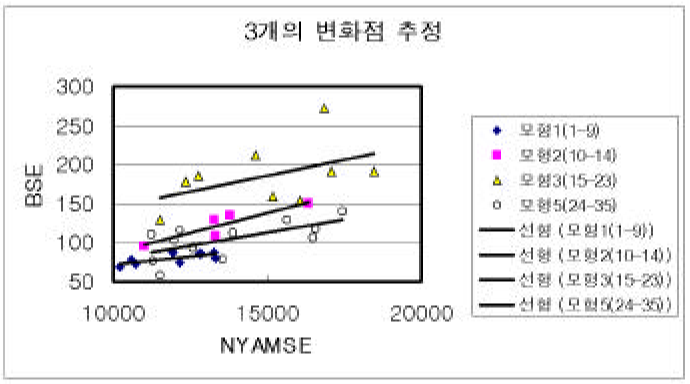

그림 4.8

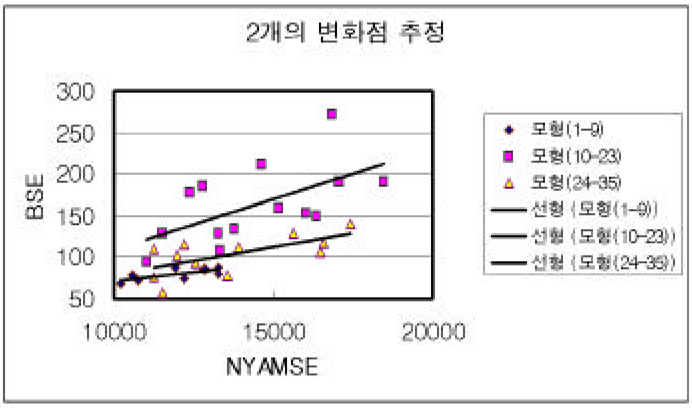

그림 4.8은 신뢰성이 의심스러운 $\hat{k}+1=20$ 인 경우를 제외하고, 변화점이 2개인 경우 $\hat{k}+1=10,24$에 대하여 3개의 회귀적선들을 추정한 것이다. 각 회귀직선을 기준으로 데이터의 산포된 분포의 정도를 가지고 분석할 때, 그림 4.4의 Chen에 의한 분석보다도 분산의 안정성을 가짐을 알 수 있다. 그러므로 제시된 TIC에 의한 변화점의 추정방법은 Chen이 제시한 SIC에 의한 방법 보다가 더 세밀하게 데이터의 집합내에서의 변화점들을 더 잘 추정함을 알 수 있다.

참고문헌

1. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In 2nd International Symposium on Information Theory*, Ed. B. N. Petrov and F. Csaki, 267-281.
2. Chen, J. (1998) "Testing for a change point in linear regression models" *Commun. statist -Theory Meth,* 27, 2481 2493
3. Chen, J. & Gupta, A.K.(1997). Testing and Locating Variance Changpoints With Application to Stock Prices, *Journal of the American Statistical Association*, 92, 739-747.
4. Schwarz, G. (1978). "Estimating the dimension of a model," *Annals of Statistics*, 6, 461-464.

[2002년 9월 접수, 2002년 9월 채택]