Journal of Statistical
Theory & Methods
1998, Vol. 9, No. 1, pp. 1 ~ 9

변화시점이 있는 영과잉-포아송모형 [1]

김 경 무 [2]

요약

영과잉-포아송모형에서 변화시점이 있는 경우, 우도비 검정통계량을 이용하여 변 화시점의 유 · 무에 대한 가설을 검정하였다. 또한 적률 및 최우추정법을 이용하여 변화시점과 몇가지 흥미있는 모수들을 추정하여 보았다. 이들 추정량을 비교하기 위하여 경험적인 평균제곱오차를 이용하였다. 변화시점이 있는 영과잉-포아송 모형과 변화시점이 없는 포아송 모형의 실례를 자료를 중심으로 설명하였다.

주제어: 변화시점, 영과잉-포아송모형, 우도비검정

1. 서 론

영과잉-포아송분포(Zero-Inflated Poisson: 이후 *ZIP*로 표기함)라 함은 이산형 확률분포 에 있어서 정상적인 포아송 확률분포보다 영의 값이 과잉 관측되는 경우를 묘사하기 위하여 원래의 확률분포를 변형시킨 것을 말한다. 포아송분포는 생산공정단계에서 발생하는 불량품 의 수에 관한 확률분포로서 지금까지 중요한 분포로 이용되어 왔다. 그러나 문명의 발달과 제품을 만들어 내는 기술의 고급화로 인하여 불량률은 현저하게 감소되어 가고 있다. 예를 들면, 반도체 분야에서 컴퓨터에 내장되어 있는 메모리 칩들은 생산공정의 정확도와 기술의 발달로 불량품으로 판정되는 경우가 드물다. 다시말하면 여러번 실시한 표본검사 중 단위당 불량품의 수는 대부분 영이고 극히 적은 경우에 한하여 불량품이 발생한다고 볼 수 있다. 이러한 경우, 기존의 포아송분포에 적합시켜 통계적인 추정 및 검정을 한다면 이는 제 3종 의 통계적 오류를 범하는 결과를 초래할 것이다.

생산공정과정에서 단위당 나타나는 불량품 수가 이러한 *ZIP* 분포를 따른다고 할 때, 불량품 수가 어떤 시점 이후 변화가 있다고 하자. 이때 미지의 시점을 변화시점(change-point)이라 한다. 변화시점에 관한 추정 및 검정 그리고 불량률의 변화가 어느 정도인 지를 추정하는 것은 중요한 일일 것이다. 이러한 *ZIP* 분포는 처음으로 Singh(1963)에 의해 소개되었

[1]이 논문은 1998학년도 대구대학교 학술연구비 지원에 의한 논문임.
[2]대구대학교 자연과학대학 통계학과 교수, (712-714) 경북 경산시 진량읍 내리리 15

다. 그러나 이 분포는 수학적인 모형으로만 인식되어 응용분야가 다양하지 못하였다. 그 이후 Lambert(1992)는 *ZIP* 회귀모형을 통하여 공변량(covariate)들의 효과를 연구하고 이를 실제의 자료에 적용하였다. 2절에서는 변화시점이 있는 *ZIP*분포를 소개하고 변화시점의 유 · 무에 대한 가설을 우도 비검정을 통하여 알아본다. 또한 흥미있는 모수들에 대한 추정량을 적률 및 최우추정법에 의해 유도하고 이들을 평균제곱오차를 이용하여 비교해 본다. 다음 3절에서는 실제 자료에 서 변화시점과 여러 흥미있는 모수들을 추정한다.

2. 변화시점이 있는 *ZIP*모형

확률변수 Y는 생산공정에서 일정 단위당 불량품이 나타나는 수로서 *ZIP* 분포를 따른다. *ZIP*는 포아송분포와 베르누이분포와의 혼합모형으로 볼 수 있다. 즉,

$$\begin{aligned} Y \quad &\sim 0, && p\text{의확률로} \\ &\sim Poisson(\lambda), && 1-p\text{의 확률,} \end{aligned}$$

여기에서 $0 \le p \le 1$ 는 불량품이 전혀 나타나지 않는 상태(perfect state)의 확률이며 $\lambda > 0$는 포아송분포의 평균이다. 이때 확률질량함수(pmf)는 아래와 같이 된다.

$$\begin{aligned} P(Y=k) &= p+(1-p)e^{-\lambda}, && k=0 \\ &= (1-p)\lambda^k e^{-\lambda}/k!, && k=1,2,\cdots \end{aligned}$$

앞으로 위 분포를 $ZIP(p,\lambda)$로 표기하기로 한다.

서로 독립인 확률변수 $Y_1, Y_2, \cdots Y_n$가 시간의 흐름에 따라 연속적으로 얻을 수 있는 관측자료라 하자. 이때 생산공정과정 중 여러가지의 원인들로 인하여 미지의 시점 c 이후 분포의 변화가 있는 다음과 같은 모형을 생각할 수 있다.

$$\begin{aligned} Y_1, Y_2, \cdots, Y_c \quad &\sim \quad ZIP\ (p,\lambda) \\ Y_{c+1}, \cdots, Y_n \quad &\sim \quad ZIP\ (p^*,\lambda^*), \end{aligned}$$

여기에서 양의 정수c는 미지의 변화시점(changepoint)이고 p^*와 λ^*는 각각 변화시점 이후 불량품이 전혀 나타나지 않는 상태의 확률과 불량품 수에 대한 평균을 의미한다. 만약 $p > p^*$ 혹은 $\lambda < \lambda^*$이 된다면, 변화시점 이후 불량품 수가 증가될 것이다. 우리는 위 모형을 변화시점이 있는 *ZIP*모형이라 하겠고 5개의 모수를 포함하고 있다.

3. 우도비 검정

생산공정과정 중 미지의 변화시점 이후 불량품 수의 변화가 있는 지를 검정하기 위하여, 변화가 없다는 귀무가설, H_0 : $p = p^*$ 그리고 $\lambda = \lambda^*$ 그리고 불량품의 수에 변화가 있다는 대립가설, H_1 : $p \neq p^*$ 혹은 $\lambda \neq \lambda^*$ 을 설정할 수 있다. 위 경우 귀무가설에 대한 로그-우도함수(log-likelihood function)는

$$l(y\,;p,\lambda) = \sum_{i=1}^{n} \ln[\{p + (1-p)e^{-\lambda}\ \}I\ (y_i = 0) + \{(1-p)\lambda^{y_i}e^{-\lambda}/y_i!\}I\ (y_i > 0)]$$

이 된다. 여기에서 $I(.)$는 지시함수(indicator function)이다. 또한 대립가설에 대한 로그-우도함수는 다음과 같다.

$$\begin{aligned} & l\ (y\ ;\ p,\lambda,c,p^*,\lambda^*) \\ &= \sum_{i=1}^{n} \ln\Big[\{p + (1-p)e^{-\lambda}\}\ I\ (i \le c,\ y_i = 0) + \{(1-p)\lambda^{y_i}e^{-\lambda}/y_i\ !\}\ \ I\ (i \le c,\ y_i > 0) \\ &+\{p^* + (1-p^*)e^{-\lambda^*}\ \}\ I\ \ (i > c,\ y_i = 0) + \{(1-p^*)\lambda^{*y_i}e^{-\lambda^*}\ /\ y_i\ !\ \}\ I\ \ (i > c,\ y_i > 0)\ \Big] \end{aligned}$$

위 대립가설에 대한 로그-우도함수가 5개의 모수로 이루어진 복잡한 함수이다. 그러므로 $p, \lambda,\ c,\ p^*$ 그리고 λ^*의 최우추정량(MLE)을 각각 $\tilde{p},\ \tilde{\lambda},\ \tilde{c},\ \tilde{p^*}$, 그리고 $\tilde{\lambda^*}$라 한다면, 이들 추정량은 간단한 형태로(closed form) 나타나지 않아 우도함수를 최대로하는 수치해석적 방법을 이용하여 구할 것이다. 연구자는 함수의 최대치를 구하는 한 가지 방법으로 Powell(1964) 방법을 이용하여 최우추정치를 구하였다. 이에 따른 우도비 검정통계량은

$$T_n = -2\{l\ (y\ ;\tilde{p},\tilde{\lambda}) - l\ (y\ ;\tilde{p},\tilde{\lambda},\tilde{c},\tilde{p^*},\tilde{\lambda^*})\}$$

이다. 위 통계량은 잘 알려져 있드시 대표본에서 자유도 3인 카이제곱분포에 접근한다. 그러므로 T_n의 값이 크면 귀무가설을 기각시킬 수 있다.

4. 추 정

변화시점이 있는 ZIP모형에서 $p,\ \lambda,\ p^*,\ \lambda^*$ 그리고 변화시점 c를 추정하는 일은 흥미있는 일일 것이다. 또한 변화시점 이후와 이전 불량품이 전혀 나타나지 않는 확률의 차이, 즉 $P(Y_i = 0|i \le c) - P(Y_i = 0|i > c)$의 추정에 관심이 있다. 변화시점을 제외한 모든 모수들의 추정방법은 적률을 이용한 추정법(MME)과 최우추정법(MLE)을 이용할 것이다. MME는 비교적 알기 쉬운 형태로 유도될 수 있으며 이를 MLE를 수치해석적(반복법)으로 구하기 위한 초기값으로 이용할 수 있다. 또한 이들 두 추정방법을 비교하기 위하여 평균제곱오

차(MSE)를 이용한다. 먼저 변화시점 c의 추정량은 변화시점 이전과 이후의 제곱합의 합이 최소가 되도록 구한다. 즉, 식 (1)을 최소로 하는 양의 정수 j $(1 \le j < n)$값을 $\hat{c}$로 생각하고 이를 변화시점에 대한 최소제곱추정량(LSE)으로 설정하였다.

$$\sum_{i=1}^{j}(x_i - \sum_{i=1}^{j} x_i/j \quad)^2 + \sum_{i=j+1}^{n} \left\{ x_i - \sum_{i=j+1}^{n} x_i/(n-j) \right\}^2 \tag{1}$$

위 변화시점에 대한 추정량 역시 식 (1)에서 직접 유도될 수 없고 주어진 표본을 이용해야 할 것이다. 이제 변화시점의 추정량이 결정이 되면, *ZIP*모형의 표본 중 변화시점 이전 표본과 이후 표본들을 대상으로 적률방법(method of moments)을 이용하면 식 (2)를 얻을 수 있다. 식 (2)의 좌변은 변화시점이 있는 *ZIP*모형에서 구한 것이고 우변은 표본에서 얻은 통계량이다. 순서대로 변화시점 이전의 평균, 제곱의 평균 그리고 변화시점 이후의 평균, 제곱평균이다.

$$\begin{aligned}
(1-p)\lambda &= \sum_{i=1}^{\hat{c}} x_i \ /\hat{c}, \\
(1-p)\lambda(1+\lambda) &= \sum_{i=1}^{\hat{c}} x_i^2 \ \ /\hat{c}, \\
(1-p^*)\lambda^* &= \sum_{i=\hat{c}+1}^{n} x_i \ \ /(n-\hat{c}), \\
(1-p^*)\lambda^*(1+\lambda^*) &= \sum_{i=\hat{c}+1}^{n} x_i^2 \ \ /(n-\hat{c} \ ).
\end{aligned} \tag{2}$$

또한 식 (2)를 모수에 관하여 정리하면, 변화시점을 제외한 나머지 모수들의 MME는 식 (3)과 같다.

$$\begin{aligned}
\hat{\lambda} \quad &= \frac{\sum_{i=1}^{\hat{c}} x_i^2}{\sum_{i=1}^{\hat{c}} x_i} - 1, \\
\hat{p} \quad &= 1 - \frac{\sum_{i=1}^{\hat{c}} x_i \ /\hat{c}}{\hat{\lambda}}, \\
\hat{\lambda^*} \quad &= \frac{\sum_{i=\hat{c}+1}^{n} x_i^2}{\sum_{i=\hat{c}+1}^{n} x_i} - 1, \\
\hat{p^*} \quad &= 1 - \frac{\sum_{i=\hat{c}+1}^{n} x_i \ \ /(n-\hat{c} \ )}{\hat{\lambda}^*}.
\end{aligned} \tag{3}$$

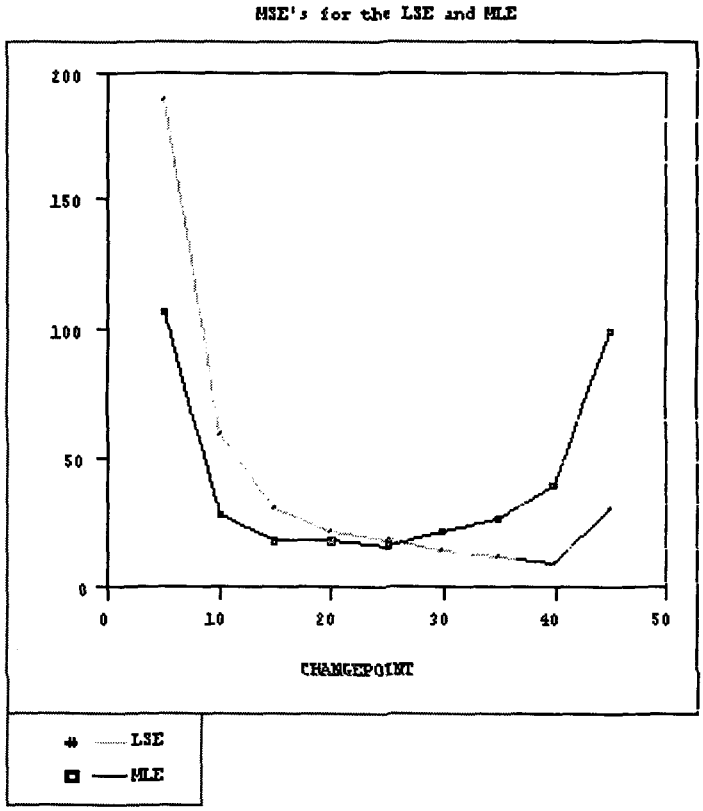
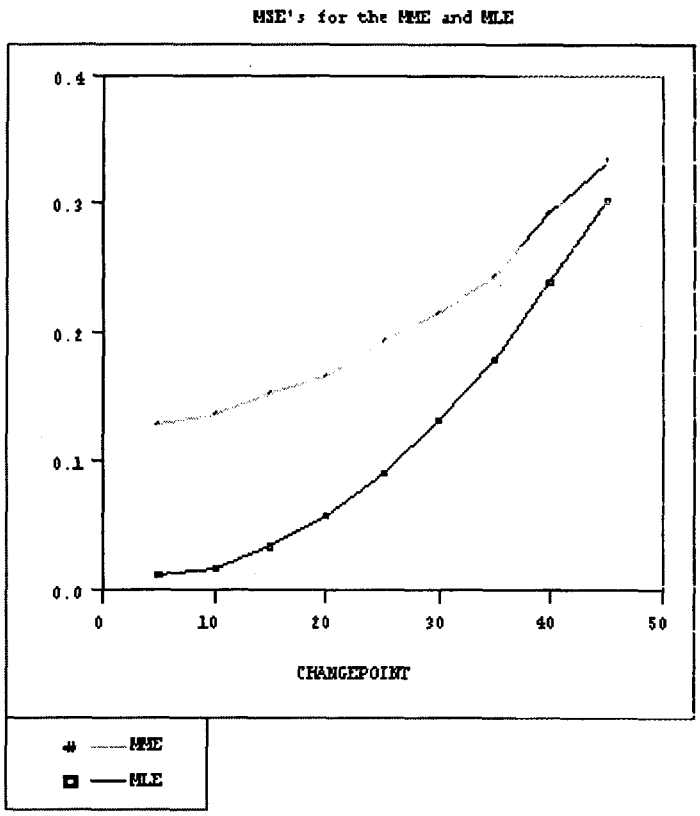


그림 4.1 MSE에 의한 변화시점추정량의 비교

물론 위 모든 추정량들은 양의 값을 지녀야 한다. 그리고 MME 추정치는 MLE를 구하기 위한 초기값으로 활용되어질 수 있다. 또한 변화시점 이전과 이후 불량품이 전혀 나타나지 않는 확률의 차이, 즉 $\Delta(p,\lambda,c,p^*,\lambda^*) = P(Y_i = 0|i \le c) - P(Y_i = 0|i > c)$의 추정량은 각 모수의 추정량에 MME를 대입하면 식 (4)를 얻을 수 있다.

$$\widehat{\Delta}(\widehat{p},\widehat{\lambda},\widehat{c},\widehat{p^*},\widehat{\lambda^*}) = \widehat{p} - (1-\widehat{p})\exp(-\widehat{\lambda}) - \widehat{p^*} + (1-\widehat{p^*})\exp(-\widehat{\lambda^*}) \quad (4)$$

변화시점 c의 LSE와 MLE를 비교하기 위하여 경험적인 MSE를 이용하였다. 많은 조건에서 모두를 비교하기가 어렵기 때문에 그 중 하나만 설명하기로 한다. 실제의 변화시점이 $5, 10, \cdots, 45$일 때 변화시점 이전과 이후의 ZIP분포의 모수들을 $p = 0.7, \lambda = 1.0, p^* = 0.2$ 그리고 $\lambda^* = 3.0$으로 하고 표본크기 50개의 난수를 추출하였다. 난수추출 방법은 실제의 변화시점 이전에는 $ZIP(p = 0.7, \lambda = 1.0)$의 난수를, 그리고 변화시점 이후는 $ZIP(p^* = 0.2, \lambda^* = 3.0)$의 난수를 추출한다. 먼저 식 (1)을 이용하여 변화시점의 LSE를 구하고 난후 나머지 모수의 MME를 구한다. 다음으로 이들의 추정치를 MLE를 구하기 위한 Powell 방법의 초기값으로 활용한다. 이러한 작업을 1,000번 반복하여 변화시점의 추정치를 구하고 이들의 MSE를 경험적으로 구하여 [그림 4.1]을 얻었다.



[그림 4.2] MSE에 의한 변화시점추정량의 비교

[그림 4.1]의 횡축은 위에서 주어진 실제의 변화시점이고, 종축은 변화시점의MLE 및 LSE의 MSE를 나타낸다. 이 그림에서 알 수 있드시 변화시점에 대한 LSE 및 MLE 모두는 변화시점이 표본의 중앙에 있을수록 MSE 값이 작게 나타난다. 이는 변화시점 이전과 이후의 표본수가 비슷할 경우 변화시점을 찾기가 쉽기 때문일 것이다. 또한 실제의 변화시점이 표본의 중앙에 위치할 경우에는 MLE 및 LSE의 MSE 값은 유사하게 나타난다. 그러나 실제의 변화시점이 표본의 앞쪽에 있을 경우는 MLE가, 표본의 뒷쪽에 있을 경우에는 MME가 바람직한 추정량으로 나타났다. 그러나 이러한 결과를 이론적으로 입증하기에는 많은 어려움이 따른다. 또한 앞으로 연구될 부분이라 생각된다. 한편 [그림 4.2]에서는 변화시점 이전과 이후 불량품이 전혀 나타나지 않는 확률의 차이, 즉 $\Delta = P(Y_i = 0|i > c) - P(Y_i = 0|i \leq c)$ 대한 MME 및 MLE의 MSE 값들을 비교해주고 있다. 이 그림 역시 표본크기 50, $p = 0.7, \lambda = 1.0, p^* = 0.2$ 그리고 $\lambda^* = 3.0$의 같은 조건에서 난수를 추출하고 이것을 이용하여 변화시점 추정치를 구하였다. 또한 변화시점 추정치를 이용하여 변화시점 이전의 모수, p, λ 그리고 변화시점 이후의 모수, p^*, λ^*의 추정치를 구하였다. 마지막으로 변화시점 이전과 이후의 불량률의 변화, 즉 $\widehat{\Delta}$는 앞에서 구한 모수들의 추정치를 대입해서 구한다. 이러한

방법을 1,000번 반복하여 모수들의 MME 및 MLE의 MSE를 얻었다.이 그림에서는 MLE가 MME보다 바람직한 추정량으로 나타났다.

5. 응용사례

<표 4.1>은 미국, 북 노스캐롤라이나주 랄리에 소재한 Nortel 회사의 불량품수에 관한 자료이다. 회사의 보안문제로 구체적인 변수의 내용은 밝힐 수 없다. 데이터 I과 II는 각각 72개의 자료 중 영이 83% 그리고 58%를 차지하고 있다. 이 자료를 변화시점이 있는 *ZIP* 모형에 적용하여 변화시점의 유무에 대한 검정과 관심있는 모수들을 추정하였다. 데이터 I은 영이 대부분을 차지하고 있으므로 기존의 포아송분포에 적용하는 것은 적당치 못하다. 데이터 I에서 변화시점이 존재하는 지를 검정하기 위하여 3절의 우도비검정을 실시하였다. 검정통계량은 $T_n = 8.94$이고 이에 따른 카이제곱분포를 이용한 p-값은 0.030이다. 한편 < 표 4.2 >는 모수의 추정값을 나타내는데, 여기에서 불량품이 전혀 나타나지 않는 상태의 확률, 즉 $\tilde{p} = 0.961, \widehat{p} = 0.774$로 추정되어 기존의 포아송분포는 적합치 않음을 알 수 있다. 그러므로 데이터 I은 변화시점이 있는 *ZIP* 모형에 적합하고 변화시점의 추정값은 $\tilde{c} = 32$이다. 또한 LSE의 값, $\widehat{c} = 32$이 되어 MLE와 LSE가 같게 나타난다. 그러나 데이터 II에서는 검정통계량은 $T_n = 5.22$이고 이에 따른 p-값은 0.156이다. 또한 <표 4.2>에서 $\tilde{p} = 0.001, \widehat{p} = 0.048$이므로 데이터 II는 변화시점이 없는 포아송분포에 적합하다고 볼 수 있다.

데이터 I이 포아송분포를 따르는 지를 알기 위하여, 카이제곱의 적합도검정을 실시하였다. 이때 카이제곱 통계량 값은 75.80, 이에 따른 자유도 4의 카이제곱 p-값은 0.000으로 데이터 I이 포아송분포를 따른다는 귀무가설을 기각하게 된다. 그러나 데이터 II에서는 카이제곱 통계량 값이 2.37로서 자유도 2의 카이제곱 p-값이 0.305가 되어 데이터 II이 포아송분포를 따른다는 귀무가설을 채택하게 된다. 이 결과는 <표 4.2>에서 데이터 I과 II의 p에 대한 추정치의 값과 같은 맥락을 이룬다.

<표 4.1> 불량품수

데이터 I
0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
4, 1, 2, 3, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 3, 0, 0, 0, 3, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 1, 0, 0, 0
데이터 II
2, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1,
1, 0, 0, 0, 2, 2, 0, 2, 0, 0, 0, 0, 2, 1, 0, 2, 0, 0, 0, 2, 0, 1, 0, 1, 1, 0, 1, 2, 1, 1, 0, 1,
2, 0, 0, 1, 2, 0, 0, 0

<표 4.2> 데이터 I과 II에 대한 모수들의 추정치

데이터	추정량	c	p	λ	p^*	λ^*	Δ
I	MME	32	0.774	1.479	0.655	1.523	-0.166
	MLE	32	0.961	1.637	0.642	1.467	0.108
II	MME	36	0.048	0.592	0.000	0.666	-0.416
	MLE	36	0.001	0.361	0.140	0.871	0.008

References

1. Lambert, D. (1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing, *Technometrics*, 34, pp. 1-14.

2. Powell, M.J.D. (1964). An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Derivatives, *Computer Jouranl*, 7, pp. 155-162.

3. Singh, S.N. (1963). A Note on Inflated Poisson Distribution, *Journal of the Indian Statistical Association*, 1, pp. 140-144.

Zero-Inflated Poisson Model with a Change-point

Kyungmoo Kim [3]

Abstract

In case of Zero-Inflated Poisson model with a change-point, likelihood ratio test statistic was used for testing hypothesis for a change-point. A change-point and several interesting parameters were estimated by using the method of moments and maximum likelihood. In order to compare the estimators, empirical mean-square-error was used. Real data for the Zero-Inflated Poisson model with a change-point and Poisson model without a change-point were examined.

[3]Professor, Department of Statistics, Taegu University, Kyungpook, 712-714 Korea.