응용통계연구 제 11권 1호, 1998년, pp. 163-175

위험률의 변화점에 대한 비모수적 추정 *

정광모 †

요 약

위험률 변화점모형에서 특별한 함수형이나 분포함수에 대한 가정을 하지 않는 일반적인 모형을 고려하였다. 이러한 모형은 지금까지 주로 다루어 왔던 상수항 위험률의 변화점모형뿐만 아니라 여러 유형의 변화점모형을 내포한다. 중도절단된 자료하에서 위험률 변화점에 관한 모수적 모형을 가정하지 않고 변화점 이전과 이후의 넬슨(Nelson) 누적위험함수 추정량의 기울기 차를 이용하여 추정량을 제안하고, 그의 점근적 성질을 규명한다. 붓스트랩 추정량의 일치성과 점근분포를 유도하고, 몇가지 분포함수의 경우에 몬테칼로 모의실험을 통해 제안된 방법의 경험적 성질을 살펴 보았다. 또한, 심장병 이식환자의 생존시간 자료를 통해 변화점을 추정하고 추정량의 붓스트랩분포를 구하였다.

1. 서론

시간에 의존하는 데이타에서 분포함수의 모수가 어떤 시점을 기준으로 변화되는 경우에 미지의 변화점을 추정하거나 변화점 유무에 대한 가설검정을 수행한다. 이러한 통계적 추론 과정을 변화점문제(change-point problem)라 한다. 변화점문제는 연구대상 및 연구방법에 따라 매우 다각적으로 연구되어 왔다. 평균이나 분산의 변화점모형, 회귀모형의 변화점문제, 위험률(hazard rate)에 대한 변화점모형 등이 있고 연구방법에 따라서는 모수 및 비모수적 방법과 베이지안 방법 등이 있다. 변화점문제에 대한 일반적인 논의는 Page(1954), Chernoff와 Zacks(1964), Bhattacharyya와 Johnson(1968), Hinkley(1970), Pettitt(1979) 등을 참고할 수 있으며, 본 연구에서는 제품이나 환자의 수명에서 자주 대두되는 위험률에 대한 변화점문제를 논의한다.

확률밀도함수가 $f(t)$이고 분포함수가 $F(t)$일 때, 위험률 $\lambda(t)$는

$$\lambda(t) = f(t)/\{1 - F(t)\}$$

와 같이 정의된다. 또한, 누적위험함수(cumulative hazard function) 는

$$\Lambda(t) = \int_0^t \lambda(u)du = -\log\{1 - F(t)\}$$

이다. 누적위험함수 $\Lambda(t)$와 분포함수 $F(t)$간의 관계식

$$F(t) = 1 - e^{-\Lambda(t)} \tag{1.1}$$

*이 논문은 1996년도 한국학술진흥재단의 자유공모과제 연구비에 의하여 연구되었음

† (609-735) 부산시 금정구 장전동 산30, 부산대학교 통계학과 교수

이 성립한다. 위험률 변화점모형의 특별한 경우로서 어떤 상수 β_1, β_2에 대해 미지의 시점 τ("변화점"이라 불림)를 기준으로

$$\lambda(t) = \begin{cases} \frac{1}{\beta_1}, & t \le \tau \\ \frac{1}{\beta_2}, & t > \tau \end{cases} \tag{1.2}$$

와 같이 표현되는 모형은 τ 이전과 이후의 분포함수가 지수분포와 대응된다. 반면에 일모수 와이블분포(one parameter Weibull distribution)의 위험률에 대한 변화점모형은

$$\lambda(t) = \begin{cases} \frac{\gamma_1}{\beta_1} t^{\gamma_1 - 1}, & t \le \tau \\ \frac{\gamma_2}{\beta_2} t^{\gamma_2 - 1}, & t > \tau \end{cases} \tag{1.3}$$

와 같이 나타낼 수 있다.

지금까지의 연구에서는 대부분 식 (1.2)의 변화점모형을 가정하고 최우추정(maximum likelihood estimation) 및 우도비검정(likelihood ratio test)이 제안되었다. Yao(1986)는 최우추정치의 일치성(consistency) 및 점근분포(asymptotic distribution)를 논의하였고, Nguyen, Rogers and Walker(1984)는 $t_{(n-1)} \le \tau < t_{(n)}$인 경우 $\beta_2 = t_{(n)} - \tau$이고 $\tau \to t_{(n)}$이면 우도함수가 비유계(unbounded)임을 지적하였다. 이와같은 문제점을 방지하기 위해 $\tau \le t_{(n-1)}$인 조건 또는 일반적으로 적당한 구간 $[\tau_0, \tau_1]$내에서 최우추정법이 논의된다.

그러나 수명데이타에서 흔히 나타나는 중도절단(censored) 관찰치를 내포하는 경우에는 변화점의 최우추정량에 대한 일치성이 보장되지 않으므로 실제로 중도절단 데이타에 모수적 방법을 적용하는데 많은 제약이 따른다. 변화점을 전후하여 동일한 분포함수간에 모수값의 변화가 일어나는 경우외에도 서로 다른 분포함수로 변화되는 상황을 자주 접하게된다. 예를들면, 변화점 이전에는 지수분포를 따르다가 변화점 이후에는 와이블분포를 따르는 변화점모형이 이에 해당된다. 이러한 데이타에 대한 변화점 문제에는 특별한 모수형을 가정하지 않는 일반적인 비모수적 변화점모형이 바람직하다. 중도절단형태는 임의중도절단(random censoring), 제1종(Type I) 및 제2종 (Type II) 중도절단 등으로 구분된다. 제1종 중도절단은 임의중도절단의 특별한 경우로 간주될 수 있으므로 여기서는 임의중도절단을 가정한다.

넬슨(Nelson) 추정량은 중도절단 데이터의 누적위험함수를 추정하는데 있어 매우 우수한 것으로 알려져 있다. 본 연구에서는 중도절단 데이터의 위험률을 특별한 형태로 가정하지 않고 변화점 이전과 이후의 넬슨 누적위험함수 추정량의 기울기 차를 이용하여 비모수적 변화점추정량을 제안하며, 추정량의 점근적 성질과 극한분포에 대해 논의하고자 한다.

2. 변화점모형과 추정

위험률에 대한 변화점모형 (1.2)와 (1.3)을 일반화하면 특별한 함수형을 가정하지 않는 $\lambda_1(t)$과 $\lambda_2(t)$를 써서

$$\lambda(t) = \begin{cases} \lambda_1(t), & t \le \tau \\ \lambda_2(t), & t > \tau \end{cases} \tag{2.1}$$

와 같이 나타낼 수 있다. 단, $\lambda_1(.) \neq \lambda_2(.)$ 이다. 변화점 모형 (2.1)은 지수분포 및 와이블분포 뿐만 아니라 그밖에 여러분포의 변화점모형을 포함하는 일반적인 모형이다. 기호 X_i를 i번째 개체의 수명, C_i를 임의중도절단 시간이라 하면 관찰시간은 X_i와 C_i의 최소값, 즉, $T_i = X_i \wedge C_i$가 되고, 임의중도절단 데이터는 $(T_i, \delta_i), i = 1, 2, \cdots, n$, 와 같이 표현된다. δ_i는 중도절단된 경우 0, 그렇지 않은 경우 1을 갖는다. 즉, $\delta_i = I_{\{X_i \leq C_i\}}$이다. 관찰치 t_i의 순서통계량을 $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$와 같이 나타내고 이에 대응되는 δ_i를 $\delta_{(i)}$ 라 하자. 또한, X_i 및 C_i에 대응되는 분포함수를 각각 F, G 라 하자.

중도절단 데이타의 경우 누적위험함수 $\Lambda(t)$에 대한 넬슨 추정량은 다음식

$$\Lambda_n(t) = \sum_{t_{(i)} \leq t} \frac{\delta_{(i)}}{n - i + 1} \tag{2.2}$$

와 같이 정의된다.

2.1. 비모수적 변화점 추정량

Chang, Chen과 Hsiung(1994)은 변화점모형 (1.2)를 가정하고 비모수적 추정량 및 점근분포에 대해 논의하였다. 일반적인 변화점모형 (2.1)에 대해 이와같은 내용을 유사한 방법으로 확장할 수 있다. 고정된 시점 t를 전후하여 넬슨추정량의 기울기 차를 이용해서 다음 통계량

$$D_n(t) = \omega(t) \left\{ \frac{\Lambda_n(\eta) - \Lambda_n(t)}{\eta - t} - \frac{\Lambda_n(t) - \Lambda_n(0)}{t} \right\} \tag{2.3}$$

을 정의하자. 단, $\omega(t) = t^p$, $0 \leq p \leq 1$, 는 가중치이고, η은 데이터의 최대값 또는 $P(T_i > \eta) \approx 0.01$ (Chang, et. al.(1994))와 같이 결정된다. 이 때, t는 비중도절단(uncensored) 관찰치 t_i'들의 집합 $T_u = \{t_{(1)}', \cdots, t_{(m)}'\}$ 에서 변화되는 것으로 가정하자. m은 비중도절단 관찰치의 갯수를 나타낸다. 넬슨추정량은 비중도절단된 관찰값에서만 증분을 갖는 계단함수(step function)이고, $p \geq 1/2$인 경우 구간 $[t'_{(i-1)}, t'_{(i)})$에서 증가함수이다. 따라서 변화점 τ의 비모수적 점추정량은

$$\hat{\tau}_n = \arg\max_{t \in T_u} |D_n(t)| \tag{2.4}$$

와 같다. 기호 $\arg\max_{t \in T_u} |D_n(t)|$는 $|D_n(t)|$를 최대로 하는 t 값을 말한다. Carlstein(1988)과 Dumbgen(1991)은 두 분포함수의 차를 이용한 비모수적 변화점 추정량을 제안한 바 있다.

2.2. 변화점 추정량의 점근적성질

변화점 추정 및 마팅게일(martingale)에 대한 기본적인 성질을 살펴보자. 기호 $N_i(t) = I_{[X_i, \infty)}(t \wedge C_i)$, $R_i(t) = I_{(0, T_i]}(t)$와 같이 정의하고 $N(t) = \sum_i N_i(t)$, $R(t) = \sum_i R_i(t)$라 하자. 식 (2.2)에 정의된 $\Lambda_n(t)$는

$$\Lambda_n(t) = \int_0^t R(s)^{-1} dN(s) \tag{2.5}$$

와 같이 표현될 수 있다. 마팅게일을 이용하여 추정량의 점근적 성질을 유도하기 위해 $M_n(t)$를

$$M_n(t) = N(t) - \int_0^t R(s)\lambda(s)ds$$

와 같이 놓으면, $M_n(t)$는 기본적인 마팅게일을 이룬다. 이 때, 다음식

$$\Lambda_n(t \wedge t_{(n)}) - \Lambda(t \wedge t_{(n)}) = \int_0^{t \wedge t_{(n)}} R(s)^{-1} dM_n(s)$$

는 평균 0이고 제곱적분가능(square integrable)한 마팅게일이 된다. Chang, et. al.(1994)은 식 (2.4)에 정의된 비모수적 변화점 추정량의 일치성 및 그 근사분포를 유도하였는데 주요 결과를 요약하면 다음 정리와 같다.

보조정리 2.1 가정된 변화점모형에 대해 변화점추정량 $\hat{\tau}_n$는 다음 성질을 만족한다.

$$\hat{\tau}_n - \tau = O_p(n^{-1})$$

보조정리 2.2 $p \geq 1/2$인 경우 적당한 조건하에서

$$n(\hat{\tau}_n - \tau) \xrightarrow{d} T_0$$

단, 기호 $\xrightarrow{d}$는 분포수렴을 나타내고, T_0는 어떤 극한과정(limit process)을 최대로 하는 확률변수이다.

3. 붓스트랩 분포

중도절단 데이터 (T_i, δ_i), $i = 1, 2, \cdots, n$, 에 대해 붓스트랩 재표집 (resampling) 절차를 간단히 설명해 보자. Akritas(1986)는 중도절단 데이터에서 카플란-마이어 (Kaplan-Meier) 분포함수 추정량의 붓스트랩 방법을 Efron(1981)과 Reid(1981)의 붓스트랩 재표집 방법에 따라 각각 구분하고 그 근사분포에 관해 논의하였다. Efron(1981)의 붓스트랩 재표집 방법은 우선 X_i의 분포함수의 추정치 F_n과 중도절단시간 C_i의 분포함수 G_n을 구한 후에 F_n으로 부터 표본 $X_1^*, \cdots, X_n^*$을 복원추출하고, 같은 요령으로 이와 독립적으로 G_n으로 부터 표본 $C_1^*, \cdots, C_n^*$를 복원추출한다. 이 때, 중도절단 붓스트랩 표본은

$$T_i^* = \min\{X_i^*, C_i^*\}, \quad \delta_i^* = I_{\{T_i^* = X_i^*\}}, \quad i = 1, \cdots, n$$

와 같이 정의된다. 특히, Efron(1981)에 의하면 F_n, G_n이 카플란-마이어 추정량인 경우 앞에서 소개된 방법에 의한 붓스트랩 표본은 중도절단 데이터 (T_i, δ_i), $i = 1, 2, \cdots, n$, 에서 복원추출하는 것과 동일하다. 여기서, F_n은 넬슨 추정량 Λ_n으로 부터 관계식 (1.2)에 의해 구해진다. 같은 방법으로 G_n을 구할 수 있다. 앞에서 정의된 붓스트랩 과정을 B번 반복하여 얻어진 붓스트랩 표본들로부터 변화점 추정치 $\hat{\tau}^{*(b)}, b = 1, \cdots, B$, 를 구하고, 추정치들의 분포로부터 평균제곱오차(mean squared error: MSE)와 백분위수(percentile)를 얻게 된

다. 반복횟수 B의 크기에 대한 일정한 규칙은 없지만 계산 시간 및 추정 대상에 따라 경험적으로 결정된다. 이에 대한 좀더 자세한 언급은 Efron과 Tibshirani(1993)에 잘 나와 있다.

Aalen(1978)은 확률적 적분(stochastic integral)을 이용하여 넬슨 위험함수 추정량을 일반화한 경험확률과정(empirical process)의 성질에 관해 논의하였다. 이러한 결과로부터, 특히 누적위험함수 추정량의 일치성(consistency) 및 점근적 정규성(asymptotic normality)이 성립한다. 붓스트랩 표본에 근거한 수식들을 기호 $*$를 써서 표현하자. 예를 들면, $N_i^*(t) = I_{[X_i^*,\infty)}(t \wedge C_i^*)$, $R_i^*(t) = I_{(0,T_i^*)}(t)$, Λ_N^*, $D_N^*(t)$와 같이 나타낼 수 있다. 추정량의 붓스트랩 분포의 점근적 성질을 유도하기 위해 식

$$Z_n^*(x) = N^*(\tau + \frac{x}{n}) - N^*(\tau)$$

라 하면 주어진 중도절단 데이터 (T_i, δ_i), $i = 1, 2, \cdots, n$, 에 대해 조건적으로 다음 정리가 성립한다.

보조정리 3.1 $Z_n^*(x) \xrightarrow{d} Z(x)$, $Z(x)$는 포아송과정이며 $0 \le x \le a$인 경우 $Z(x)$의 강도함수는 $f(\tau+)(1-G(\tau))$, $-Z(-x)$의 강도함수는 $f(\tau-)(1-G(\tau))$이다. 또한 $\{Z(x)|x \in [0,a]\}$와 $\{Z(x)|x \in [-a,0]\}$ 는 서로 독립이다.

증명: 집합 C 를 연속함수들의 집합이라 할 때 임의의 $r(x) \in C$에 대해

$$\int_{-a}^{a} r(x) dZ_n^*(x) \xrightarrow{w} \int_{-a}^{a} r(x) dZ(x) \tag{3.1}$$

임을 보이면 된다. 여기서 기호 $\xrightarrow{w}$ 는 약수렴(weak convergence)을 나타낸다. $r(x)$가 단순함수(simple function)일 때 식 (3.1)이 성립함을 보이고 이것을 일반적인 연속함수로 확장하면 된다. 먼저 유한차원 분포를 갖는 $Z_n^*(x)$에 대해 이를 보이자. $-a \le x_2 < x_1 \le 0 \le y_1 < y_2 \le a$일 때 구간 $(\tau + \frac{x_2}{n}, \tau]$에서 어떤 개체의 수명이 끊어질 확률 p_{1n}은

$$\begin{aligned} p_{1n} &= P[\tau + \frac{x_1}{n} < X_i^* \le \tau, X_i^* \le C_i^*] \\ &= \{F_n(\tau) - F_n(\tau + \frac{x_1}{n})\}\{1 - G_n(\tau + \frac{x_1}{n})\} \\ &= \{F_n(\tau) - F_n(\tau + \frac{x_1}{n})\}\{1 - G_n(\tau) + o(\frac{1}{n})\} \end{aligned}$$

같은 방법으로 구간 $(\tau + \frac{x_2}{n}, \tau + \frac{x_1}{n}]$, $(\tau, \tau + \frac{y_1}{n}]$, $(\tau + \frac{y_1}{n}, \tau + \frac{y_2}{n}]$ 에서 수명이 끊어질 확률을 차례로 p_{2n}, p_{3n}, p_{4n}이라 하면

$$\begin{aligned} p_{2n} &= P[\tau + \frac{x_2}{n} < X_i^* \le \tau + \frac{x_1}{n}, X_i^* \le C_i^*] \\ &= \{F_n(\tau + \frac{x_1}{n}) - F_n(\tau + \frac{x_2}{n})\}\{1 - G_n(\tau) + o(\frac{1}{n})\}. \end{aligned}$$

또한,

$$\begin{aligned} p_{3n} &= P[\tau < X_i^* \le \tau + \frac{y_1}{n}, X_i^* \le C_i^*] \\ &= \{F_n(\tau + \frac{y_1}{n}) - F_n(\tau)\}\{1 - G_n(\tau) + o(\frac{1}{n})\} \end{aligned}$$

이고,

$$\begin{aligned} p_{4n} &= P[\tau + \frac{y_1}{n} < X_i^* \le \tau + \frac{y_2}{n}, X_i^* \le C_i^*] \\ &= \{F_n(\tau + \frac{y_2}{n}) - F_n(\tau + \frac{y_1}{n})\}\{1 - G_n(\tau) + o(\frac{1}{n})\} \end{aligned}$$

와 같이 되므로, 다항확률분포로부터

$$\begin{aligned} & P[Z_n^*(x_1) = -b_1, Z_n^*(x_2) - Z_n^*(x_1) = -b_2, Z_n^*(y_1) = d_1, Z_n^*(y_2) = d_2] \\ = \; & p_{1n}^{b_1} p_{2n}^{b_2} p_{3n}^{d_1} p_{4n}^{d_2} p_{5n}^{n-b_1-b_2-d_1-d_2} \end{aligned}$$

단,

$$\begin{aligned} p_{5n} &= 1 - p_{1n} - p_{2n} - p_{3n} - p_{4n} \\ &= 1 - \{F_n(\tau + \frac{y_2}{n}) - F_n(\tau + \frac{x_2}{n})\}\{1 - G_n(\tau) + o(\frac{1}{n})\} \end{aligned}$$

이다. 여기서 $F_n(t)$, $G_n(t)$는 각각 $F(t)$, $G(t)$에 확률수렴하므로 포아송분포의 극한과정을 유도하는 방법과 유사한 근사공식을 이용하면 정리의 결과를 얻을 수 있다. 따라서, $Z_n(x)$가 일반적인 유한차원 분포인 경우에 정리의 결과가 성립한다. □

앞의 증명은 주어진 데이터 (T_i, δ_i), $i = 1, 2, \cdots, n$, 에 대해 조건적이며, 이와 같은 성질이 거의 모든 표본열 (for almost all sample sequences)에 대해 성립한다. $S(t)$를 T_i의 생존함수라 할 때 다음 정리를 얻는다.

보조정리 3.2

$$n\{\Lambda_n^*(\tau + \frac{x}{n}) - \Lambda_n^*(\tau)\} \xrightarrow{d} S^{-1}(\tau) Z(x)$$

증명: 관계식 (2.5)로부터

$$n\{\Lambda_n^*(\tau + \frac{x}{n}) - \Lambda_n^*(\tau)\} = S^{-1}(\tau) Z_n^*(x) + \int_\tau^{\tau + \frac{x}{n}} \{(\frac{R^*(t)}{n})^{-1} - S^{-1}(\tau)\} dN^*(t).$$

주어진 데이터 (T_i, δ_i), $i = 1, 2, \cdots, n$, 에 대해 앞식의 첫 번째 항은 보조정리 3.1에 의해 분포수렴하므로 두번째 항이 0에 확률수렴함을 보이면 된다.

$$\begin{aligned} & \sup_{[-a,a]} \left| \int_\tau^{\tau + \frac{x}{n}} \{(\frac{R^*(t)}{n})^{-1} - S^{-1}(\tau)\} dN^*(t) \right| \\ \le \; & \sup_{[\tau - \frac{a}{n}, \tau + \frac{a}{n}]} \left| \{(\frac{R^*(t)}{n})^{-1} - S^{-1}(\tau)\} \right| \{N^*(\tau + \frac{a}{n}) - N^*(\tau - \frac{a}{n})\} \\ \le \; & \left[\max\{(\frac{R^*(\tau - \frac{a}{n})}{n})^{-1}, (\frac{R^*(\tau + \frac{a}{n})}{n})^{-1}\} - S^{-1}(\tau) \right] \{N^*(\tau + \frac{a}{n}) - N^*(\tau - \frac{a}{n})\}. \end{aligned}$$

위 맨 끝식의 첫 번째 항은 0에 수렴하고 두 번째 항은 보조정리 3.1에 의해 분포수렴하므로 정리의 결과가 성립함을 알 수 있다. 앞의 사실은 거의 모든 표본 데이터열에 대해 조건적으로 성립하는 결과이다. □

정리 3.1 $p \geq \frac{1}{2}$인 경우, 거의 모든 표본열 (T_i, δ_i), $i = 1, \cdots, n$, 에 대해

$$n[D_n^*(\tau + \frac{x}{n}) - D_n^*(\tau)] \xrightarrow{d} W(x)$$

이다. 단, $W(x) = xu - vZ(x)$이고, u, v는 τ 및 η의 어떤 함수이다.

증명: $D_n^*(t)$의 정의로부터 관계식

$$D_n^*(t) = \Lambda_n^*(\eta)t^p(\eta - t)^{p-1} - \Lambda_n^*(t)\eta t^{p-1}(\eta - t)^{p-1}$$

이 성립하므로

$$\begin{aligned} n[D_n^*(\tau + \frac{x}{n}) - D_n^*(\tau)] &= n\Lambda_n^*(\eta)\{(\tau + \frac{x}{n})^p(\eta - \tau - \frac{x}{n})^{p-1} - \tau^p(\eta - \tau)^{p-1}\} \\ &- \eta n\Lambda_n^*(\tau)\{(\tau + \frac{x}{n})^{p-1}(\eta - \tau - \frac{x}{n})^{p-1} - \tau^{p-1}(\eta - \tau)^{p-1}\} \\ &- \eta n\{\Lambda_n^*(\tau + \frac{x}{n}) - \Lambda_n^*(\tau)\}(\tau + \frac{x}{n})^{p-1}(\eta - \tau - \frac{x}{n})^{p-1} \end{aligned}$$

여기서 $\frac{d}{dt}\{t^p(\eta - t)^{p-1}\} = t^{p-1}(\eta - t)^{p-2}\{p\eta - 2p\tau + \tau\}$인 관계를 이용하면, $n \to \infty$일 때 첫째 및 둘째 항은 극한값에 수렴하게 되고 이 두 값의 합은 xu에 수렴한다. 단, $u = \Lambda(\eta)\tau^{p-1}(\eta - \tau)^{p-2}\{p\eta - 2p\tau + \tau\} - \eta\Lambda(\tau)\tau^{p-2}(\eta - \tau)^{p-2}(\eta - 2\tau)(p-1)$ 이다. 또한 유사한 방법으로 보조정리 3.2로부터 세 번째 항이 $vZ(x)$에 약수렴 하는 것을 보일 수 있다. 단, $v = \eta\tau^{p-1}(\eta - \tau)^{p-1}S^{-1}(\tau)$ 이다. □

정리 3.2 $u > 0$, $p \geq \frac{1}{2}$일 때, 거의 모든 표본열 (T_i, δ_i), $i = 1, \cdots, n$, 에 대해 $n(\hat{\tau}_n^* - \hat{\tau}_n) \xrightarrow{d} T_0$ 가 성립한다. 단, T_0는 정리 3.1에서 정의된 극한과정 $W(x) = xu - vZ(x)$을 최대로 하는 x값이다.

증명: 정리 3.1과 약수렴에 대한 연속사상정리(continuous mapping theorem)에 의해 정리 내용을 유도할 수 있으며, 자세한 증명 절차는 Chang, et. al.(1994)의 증명과 유사한 방법으로 전개된다. □

따름정리 3.1 거의 모든 표본열 (T_i, δ_i), $i = 1, \cdots, n$, 에 대해 $\hat{\tau}_n^* - \hat{\tau}_n = O_p(n^{-1})$ 가 성립한다.

4. 모의실험계획 및 예제

4.1. 모의실험계획

앞에서 논의된 변화점 추정량의 점근적 성질을 검토하기 위해 중도절단 관찰치의 비율, 변화점의 위치 및 분포함수의 형태에 따라 다음과 같은 모의실험계획을 세울 수 있다. 변화점모형 (1.3)에서 모수값에 따라 다음과 같이 3가지 경우를 고려한다.

사례 1: $\beta_1 = 2.0$, $\beta_2 = 1.0$, $\gamma_1 = \gamma_2 = 1.0$, $\tau = 1.20$인 지수분포에서 지수분포로의 변화점모형

사례 2: $\beta_1 = 2.0$, $\beta_2 = 1.0$, $\gamma_1 = 1.0$, $\gamma_2 = 2.0$, $\tau = 1.20$인 지수분포에서 와이블분포로의 변화점모형

사례 3: $\beta_1 = 2.0$, $\beta_2 = 1.0$, $\gamma_1 = 2.0$, $\gamma_2 = 0.5$, $\tau = 1.10$인 와이블분포에서 와이블분포로의 변화점모형

또한, 중도절단분포는 균일분포를 가정하고 중도절단 관찰치의 비율이 20% 및 40%인 경우를 고려한다. 모의실험계획의 변화점 τ는 전체 데이터에서 변화점의 위치가 약 45 백분위점에 오도록 정해진 것이다. 그림 4.1 - 그림 4.3은 모의실험계획 사례 1 - 사례 3에서 생성된 데이터에 대한 넬슨누적위험함수 추정량을 나타내며 중도절단비율에 따라 (a) 20% (b) 40% 등으로 구분하였다.

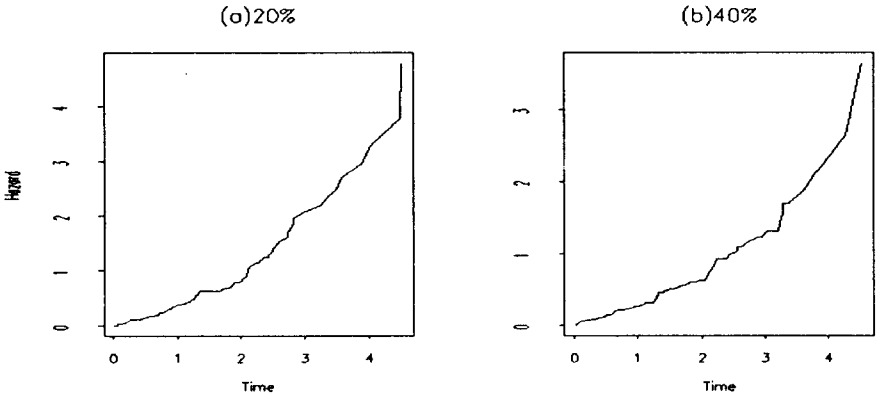

그림 4.1: 누적위험함수(사례1)

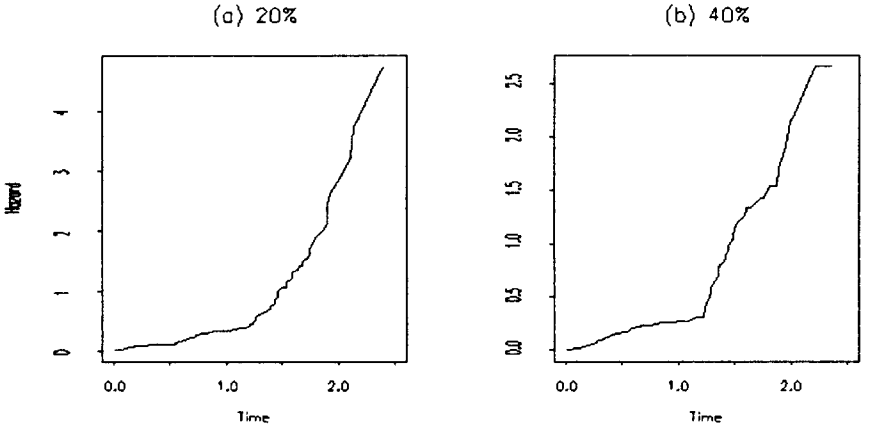

그림 4.2: 누적위험함수(사례2)

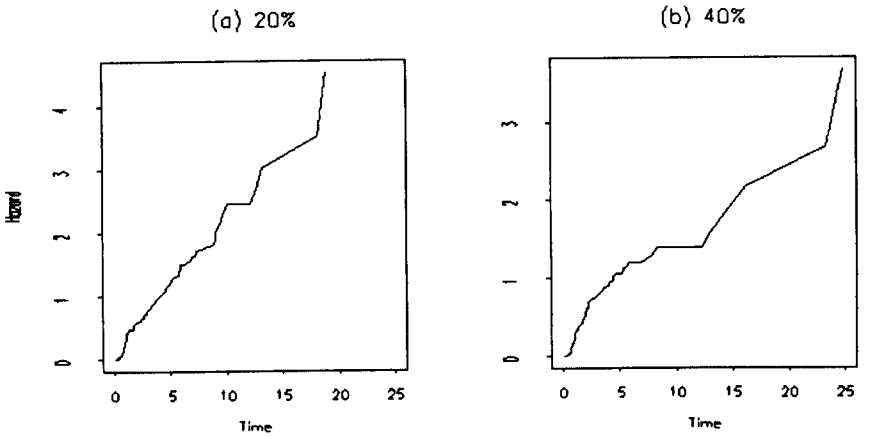

그림 4.3: 누적위험함수(사례3)

표본크기는 $n = 100$으로 하였으며 붓스트랩 반복은 3,000번을 수행하고 주어진 데이터에 대한 변화점의 비모수적 추정량 $\hat{\tau}$과 붓스트랩 추정량의 평균제곱오차(mean squared error:MSE) 및 백분위점(0.5%, 2.5%, 50%, 97.5%, 99.5%)을 표 4.1 - 표 4.3에 나타내었다.

표 4.1: $\beta_1 = 2.0$, $\beta_2 = 1.0$, $\gamma_1 = \gamma_2 = 1.0$, $\tau = 1.20$ (사례1)

			백분위수				
중도절단비율	$\hat{\tau}$	MSE	0.5%	2.5%	50%	97.5%	99.5%
20%	1.2544	.0015	1.1855	1.2001	1.2554	1.3358	1.3515
40%	1.2443	.0092	1.2443	1.2443	1.2691	1.4609	1.4609

표 4.2: $\beta_1 = 2.0$, $\beta_2 = 1.0$, $\gamma_1 = 1.0$, $\gamma_2 = 2.0$, $\tau = 1.20$ (사례2)

			백분위수				
중도절단비율	$\hat{\tau}$	MSE	0.5%	2.5%	50%	97.5%	99.5%
20%	1.3827	.0019	1.3827	1.3827	1.4152	1.4912	1.4912
40%	1.2211	.0014	1.2211	1.2211	1.2298	1.3501	1.4737

표 4.3: $\beta_1 = 2.0$, $\beta_2 = 1.0$, $\gamma_1 = 2.0$, $\gamma_2 = 0.5$, $\tau = 1.10$ (사례3)

			백분위수				
중도절단비율	$\hat{\tau}$	MSE	0.5%	2.5%	50%	97.5%	99.5%
20%	1.0588	.0151	0.9369	0.9369	1.0588	1.2312	1.2312
40%	1.1209	.0491	0.8156	0.8156	1.2990	1.4793	1.4793

상수에서 상수로의 변화점모형인 경우 추정 및 오차가 양호한 것으로 나타났으며, 와이블분포의 위험률에 대한 변화점모형을 나타낸 표 4.3에서는 MSE가 다른 모형에 비해 월등히 큰 것을 알 수 있다. 중도절단 비율에 따라서도 그 비율이 클 때 대체로 MSE가 커지고 있으나, 표 4.2의 경우는 오히려 반대 현상을 나타내는 데 생성된 모의자료의 변이성 때문에 변화점의 추정치가 주어진 값과 상대적으로 크게 벗어난 원인으로 생각된다.

4.2. 예 제

심장이식 수술을 받은 184명 환자의 수명데이타(Stanford heart transplant data: Cox and Oakes(1984))에 관한 적용 결과이다. 이들 환자중 119명은 그 후 조사기간내에 사망했다. 그림 4.4는 넬슨 누적위험함수 추정량을 보여주며, 본 연구에서 제안된 비모수적 변화점 추정량은 $\hat{\tau} = 68$ (일) 이다. 이 값은 Loader(1991)가 최우추정법에 의해 구한 68.01(일)과 근사함을 알 수 있다. 한편 이 자료에 대한 붓스트랩 근사분포는 다음 표 4.4에 주어져있다.

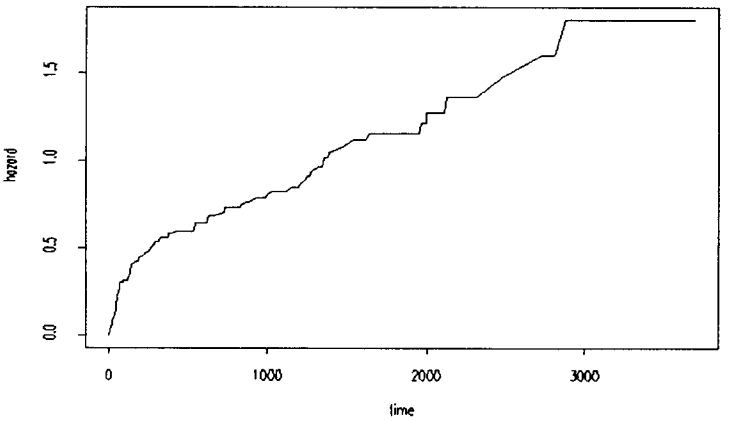

그림 4.4: 넬슨누적위험함수

표 4.4: $\hat{\tau}$의 붓스트랩 분포(B=5,000)

$\hat{\tau}$	51	54	65	66	68	148	기 타	합 계
빈도	294	60	70	509	3549	216	302	5000
비율(%)	.0588	.0120	.0140	.1018	.7098	.0432	.0604	1.0

5. 요약 및 결론

상수에서 상수로의 위험률 변화점모형을 일반화하여 특정한 함수형을 가정하지 않은 변화점모형을 고려하였다. 이러한 변화점모형은 지금까지 주로 연구 대상이었던 지수분포

의 위험률뿐만 아니라 신뢰성분야에서 많이 연구되는 와이블분포의 변화점모형을 포함한다. 더욱이 변화점을 전후하여 서로 다른 분포형에 대응되는 위험률 함수의 변화점모형을 내포한다는 점에서 매우 일반적으로 적용할 수 있는 비모수적 접근 방법이다.

본 연구에서는 붓스트랩을 이용하여 추정량의 점근분포를 유도하고 몇 가지 분포형에서 모의실험을 통해 경험적 결과를 확인하였다. 또한, 중도절단 데이터의 변화점모형에서 자주 인용되는 실제 예를 통해 그 실용성을 모색하였다. 변화점의 신뢰영역(confidence region)에 대한 연구는 본 과제에서 다루지 못했으며 앞으로 흥미있는 연구분야가 되리라고 생각된다. 아울러 좀 더 폭 넓은 모의실험 및 기존의 모수적 접근법과의 비교도 앞으로의 연구과제로 남겨둔다.

참고문헌

[1] Aalen, O.(1978), Nonparametric Inference for a Family of Counting Processes, *The Annals of Statistics*, 6(4), 701-726

[2] Akritas, M.G.(1986), Bootstrapping the Kaplan-Meier Estimator, *Journal of the American Statistical Association*. 81(396), 1032-1038,

[3] Bhattacharyya, G. K. and Johnson, R. A.(1968), Nonparametric Tests for Shift at Unknown Time Point, *Annals of Mathematical Statistics*, 39, 1731-1743

[4] Carlstein, E.(1988), Nonparametric Change-Point Estimation, *The Annals of Statistics*. 16(1), 188-197

[5] Chang, I. S., Chen, C. H. and Hsiung, C. A.(1994), Estimation in Change-Point Hazard Rate Models With Random Censorship, Change-Point Problems, 78-92, Institute of Mathematical Statistics, Lecture Notes 23

[6] Chernoff, H. and Zacks, S.(1964), Estimating the Current Mean of a Normal Distribution Which is Subject to Changes in Time, *Annals of Mathematical Statistics*, 35, 999-1018

[7] Cox, D. R. and Oakes, D.(1984), *Analysis of Survival Data*. London: Chapman and Hall

[8] Dumbgen, L.(1991), The Asymptotic Behaviour of Some Nonparametric Change-Point Estimators, *The Annals of Statistics*, 19(3), 1471-1495

[9] Efron, B.(1981), Censored Data and the Bootstrap, *Journal of the American Statistical Association*, 76, 312-19

[10] Efron, B. and Tibshirani, R. J.(1993), *Introduction to the Bootstrap*, Chapman and Hall

[11] Hinkley, D. V.(1970), Inference About the Change-Point in a Sequence of Random Variables, *Biometrika*, 57, 1-17

[12] Loader, C. R.(1991), Inference for a Hazard Rate Change Point, *Biometrika*, 78, 749-757

[13] Nguyen, H. T., Rogers, G. S. and Walker, E. A. (1984), Estimation in Change-Point Hazard Rate Models, *Biometrika*, 71, 299-304

[14] Page, E. S.(1954), Continuous Inspection Schemes, *Biometrika*, 41, 100-115

[15] Pettitt, A. N.(1979), A Non-parametric Approach to the Change-Point Problem, *Applied Statistics*, 28, 126-135

[16] Reid, N.(1981), Estimating the Median Survival Time, *Biometrika* , 68, 601-608

[17] Yao, Yi-Ching(1986), Maximum Likelihood Estimation in Hazard Rate Models with a Change-Point, *Communications in Statistics-Theory and Methods.* 15(8), 2455-2466

[1997년 8월 접수, 1997년 12월 최종수정]

Nonparametric Estimation of Hazard Rates Change-Point *

Kwang Mo Jeong †

ABSTRACT

The change of hazard rates at some unknown time point has been the interest of many statisticians. But it was restricted to the constant hazard rates which correspond to the exponential distribution. In this paper we generalize the change-point model in which any specific functional forms of hazard rates are not assumed. The assumed model includes various types of changes before and after the unknown time point. The Nelson estimator of cumulative hazard function is introduced. We estimate the change-point maximizing slope changes of Nelson estimator. Consistency and asymptotic distribution of bootstrap estimator are obtained using the martingale theory. Through a Monte Carlo study we check the performance of the proposed method. We also explain the proposed method using the Stanford Heart Transplant Data set.

*This research was supported by Non Directed Research Fund, Korea Research Foundation, 1996

† Pusan National University