Journal of the Korean Data & Information Science Society
2014, **25**(2), 393–402
http://dx.doi.org/10.7465/jkdi.2014.25.2.393
한국데이터정보과학회지

중도절단모형이 지수분포의 척도모수추정에 미치는 영향†

김남현[1]

[1]홍익대학교 기초과학과

접수 2014년 1월 28일, 수정 2014년 2월 20일, 게재확정 2014년 3월 11일

요 약

수명시간 분석에서 자주 이용되는 분포 중 하나는 지수분포이다. 본 논문에서는 임의중도절단 자료의 분석에서 중도절단모형이 지수분포의 모수추정에 어떤 영향을 주는지에 대해서 알아보았다. 고려한 중도절단모형은 Koziol-Green 모형과 일반화 지수분포 모형으로 이들은 의미상 매우 다른 모형이다. 모의실험을 통해서 살펴본 결과 중도절단모형이 모수의 평균적인 추정값에는 크게 영향을 주지 않는다고 보이나 가정한 모형이 실제의 모형과 차이가 심하게 나는 경우 추정량의 MSE가 커지는 경향을 보였다.

주요용어: 임의중도절단, 중도절단모형, 최대우도추정량, 코지올-그린 모형.

1. 서론

생존분석에서 가장 자주 이용되는 분포 중 하나는 지수분포이고, 중도절단의 형태로는 제 1종 중도절단 (type I censoring), 제 2종 중도절단 (type II censoring), 임의중도절단 (random censoring) 등을 자주 고려한다. 중도절단 자료에 대한 일반적인 사항은 Cohen (1991), Lee와 Wang (2003), Tableman과 Kim (2004) 등을 참고로 한다. 임의중도절단의 경우에는 중도절단분포에 대한 추가적인 가정이나 모형이 필요하다. 대표적인 지수분포에 대한 중도절단모형 (random censorship model)으로는 Koziol-Green 모형을 들 수 있다. 이는 결국 중도절단분포 역시 지수분포를 가정하는 것으로 직렬모형으로 생각할 수 있다. 본 논문에서는 중도절단분포에 대해서 이와 달리 병렬모형으로 해석할 수 있는 분포 또한 고려한다. 이 경우 중도절단분포는 일반화 지수분포가 된다. 일반화 지수분포는 Gupta와 Kundu (1999)에서 소개되었으며 와이블분포, 감마분포 등과 같이 수명자료의 분석에 유용하게 이용된다. 실제 자료분석 상황에서 중도절단모형을 정하는 것은 그리 쉽지는 않을 것이다. 본 논문에서는 상반된 의미를 갖는다고 볼 수 있는 각각의 모형을 가정했을 때, 이러한 가정이 지수분포의 척도모수 추정에 어떤 영향을 미치는지에 대해 생각해 보고 이를 모의실험을 통하여 알아보고자 한다.

2절에서는 구체적으로 중도절단모형을 소개하고 각각의 경우 척도모수가 어떻게 추정되는지를 언급한다. 3절에서는 모의실험결과를 소개하고 실제 자료에 적용해 본다. 4절에서는 결론을 간략하게 언급한다.

† 이 논문은 2013년도 홍익대학교 학술연구진흥비에 의하여 지원되었음.
[1] (121-791) 서울시 마포구 와우산로 94 (상수동), 홍익대학교 기초과학과, 교수.
E-mail: nhkim@hongik.ac.kr

2. 중도절단모형에 따른 추정량 비교

$Y_1^0, ..., Y_n^0$는 연속확률분포 F_{Y^0}에서의 확률표본이고, $T_1, ..., T_n$은 $Y_1^0, ..., Y_n^0$에 독립인, 연속분포 F_T에서의 중도절단 확률변수라고 하자. Y_i^0는 T_i에 의해서 우측 중도절단 (right censored)되고, 관측되는 자료는 (Y_i, δ_i), $i = 1, ..., n$이다. 여기서 $Y_i = \min(Y_i^0, T_i)$이고

$$\delta_i = \begin{cases} 1 & \text{if } \ Y_i = Y_i^0 \\ 0 & \text{if } \ Y_i = T_i \end{cases}$$

이다.

Koziol과 Green (1976)은 중도절단분포 F_T에 대해서 모형

$$1 - F_T = (1 - F_{Y^0})^\beta \tag{2.1}$$

을 소개하였다. 여기서 β는 고정된 양수이며, 이를 중도절단모수 (censoring parameter)라고 불렀다.

$$\gamma = P(Y_i^0 > T_i) = \int_{-\infty}^{\infty} (1 - F_{Y^0}(y)) dF_T(y) = \int_0^1 \beta (1-x)^\beta dx = \frac{\beta}{\beta+1} \tag{2.2}$$

로 β가 증가할수록 중도절단관측의 기대비율이 높아지기 때문이다. Csörgő와 Horváth (1981)는 식 (2.1)을 Koziol-Green 모형이라고 불렀으며, Chen 등 (1982)에서는 이 모형의 특징과 의미에 대해서 설명하고 있다. 이를 간단히 설명하면 다음과 같다. 신뢰성 모형에서 각각 수명시간이 Y^0와 T인 두 개의 성분을 가진 직렬시스템을 생각하자. 이 시스템은 두 성분이 모두 작동해야 정상적으로 작동가능하므로 이 시스템의 수명시간은 $Y = \min(Y^0, T)$이다. 만일 첫번째 성분의 수명시간이 Y^0, 수명분포 F_{Y^0}이고 두번째 성분은 β개의 부성분 (subcomponents)으로 직렬연결 되어있고 각각 독립적으로 작용하며 동일한 수명분포 F_{Y^0}를 따른다고 하자. 그러면 두번째 성분의 수명시간 T는 β개의 부성분의 수명시간 중 최소가 되므로 T의 분포는 식 (2.1)의 Koziol-Green 모형을 따르게 된다. 이 경우 β는 양의 정수일 때 의미가 있으나 일반적으로 β는 양수라고 가정한다.

만일 두번째 성분이 직렬이 아닌 병렬로 구성되어 있고, 마찬가지로 β개의 부성분을 가진다고 가정하자. 이 시스템이 제대로 작동하기 위해서는 β개 중 적어도 하나의 성분이 작동하면 된다. 따라서 분포 F_T는

$$F_T = (F_{Y^0})^\beta \tag{2.3}$$

가 된다. 이 경우에는 중도절단관측의 기대비율이

$$\gamma = P(Y_i^0 > T_i) = \frac{1}{\beta+1} \tag{2.4}$$

이다.

본 논문에서는 연속확률분포 F_{Y^0}에서의 확률표본 $Y_1^0, ..., Y_n^0$가 모수 σ를 가진 지수분포 $\mathcal{E}(1/\sigma)$를 따를 때, 즉 $F_{Y^0}(y) = 1 - e^{-y/\sigma}$, $y > 0$, $\sigma > 0$일 때를 가정하고, 중도절단 확률변수 $T_1, ..., T_n$의 분포 F_T에 대한 가정이 모수 σ의 추정에 미치는 영향에 대해 알아보고자 한다. 이 경우 식 (2.1)의 중도절단분포 F_T는 $\mathcal{E}(\beta/\sigma)$가 되고, 식 (2.3)의 F_T는

$$F_T(t) = (1 - e^{-t/\sigma})^\beta, \quad t > 0$$

이 된다. 이는 모수 β, σ를 가진 일반화 지수분포 (generalized exponential distribution; GE)이다. β를 형상모수 (shape parameter), σ를 척도모수(scale parameter)라고 하며, 모수 β, σ를 가진 일반

화 지수분포를 $GE(\beta,\sigma)$라고 쓰자. GE 분포는 Gupta와 Kundu (1999)에서 소개되었으며 와이블분포, 감마분포 등과 같이 수명자료의 분석에 유용하게 이용된다. Gupta와 Kundu (2001a, 2001b, 2002, 2007), Raqab (2002), Cho 등 (2013)은 GE 분포의 추론에 대해 좀 더 자세히 살펴보았다.

Y_i^0, T_i의 확률밀도함수를 각각 f_{Y^0}, f_T, 생존함수 (survival function)를 각각 $\overline{F}_{Y^0}$, $\overline{F}_T$라고 하면, (Y_i,δ_i), $Y_i=\min(Y_i^0,T_i)$의 우도함수 (likelihood function)는

$$\prod_{i=1}^{n}\left(f_{Y^0}(y_i)\overline{F}_T(y_i)\right)^{\delta_i}\left(f_T(y_i)\overline{F}_{Y^0}(y_i)\right)^{1-\delta_i} \tag{2.5}$$

이다. 다음 각 경우에 모수 σ의 최대우도추정량 (maximum likelihood estimator; MLE)를 생각하여 보자.

(i) $T_1,...,T_n$의 분포 F_T가 σ를 포함하지 않는 경우

이 경우는 우도함수가

$$\begin{aligned}&\prod_{i=1}^{n}\left(f_{Y^0}(y_i)\right)^{\delta_i}\left(\overline{F}_{Y^0}(y_i)\right)^{1-\delta_i}\\ =\ &\prod_{i=1}^{n}\left(1/\sigma e^{-y_i/\sigma}\right)^{\delta_i}\left(e^{-y_i/\sigma}\right)^{1-\delta_i}\\ =\ &(1/\sigma)^{n_u}e^{-\sum_{i=1}^{n}y_i/\sigma}\end{aligned}$$

이다. 여기서 n_u는 $\sum_{i=1}^{n}\delta_i$로 n개 중 중도절단되지 않은 자료의 수를 말한다. 이로부터 σ의 MLE가

$$\hat{\sigma}_0=\frac{\sum_{i=1}^{n}y_i}{n_u} \tag{2.6}$$

임을 쉽게 알 수 있다.

(ii) $1-F_T=(1-F_{Y^0})^{\beta}$, $\beta>0$인 경우

이 경우는 우도함수인 식 (2.5)가

$$\prod_{i=1}^{n}\left(1/\sigma e^{-y_i/\sigma}e^{-\beta y_i/\sigma}\right)^{\delta_i}\left(\beta/\sigma e^{-\beta y_i/\sigma}e^{-y_i/\sigma}\right)^{1-\delta_i}$$

이다. 우선 모수 β를 주어진 값으로 가정하면 σ의 MLE는

$$\hat{\sigma}_1=\frac{(1+\beta)\sum_{i=1}^{n}y_i}{n} \tag{2.7}$$

임을 쉽게 알 수 있다. 만일 β가 미지이면 β의 MLE는 $(1-n_u)/n_u$가 되므로 이로부터 σ의 MLE가 (i)의 경우와 같이 $\sum_{i=1}^{n}y_i/n_u$가 된다. 그러나 다음 모형과의 비교를 위하여 일단 β를 주어진 값으로 가정하자.

(iii) $F_T=(F_{Y^0})^{\beta}$, $\beta>0$인 경우

이 경우는 $F_T(t)$가 일반화 지수분포 $GE(\beta,\sigma)$를 따른다. 관측자료는 $Y_1,...,Y_n$의 순서통계량 $Y_{(1)}\leq\cdots\leq Y_{(n)}$에 대해서 $(Y_{(i)},\delta_{(i)})$, $1\leq i\leq n$으로도 쓸 수 있다. 여기서 $\delta_{(i)}$는 i번째 순서통계량이 중도절단 되지 않았을 때 1을 갖는다. 우선 β는 주어진 값이라고 가정하자. $(Y_{(i)},\delta_{(i)})$, $i=1,...,n$의 우도함수는

$$L(\sigma)=n!\prod_{i=1}^{n}\left(f_{Y^0}(y_{(i)})\bar{F}_T(y_{(i)})\right)^{\delta_{(i)}}\left(f_T(y_{(i)})\bar{F}_{Y^0}(y_{(i)})\right)^{1-\delta_{(i)}}$$

이므로, 로그우도함수 $l(\sigma) = \ln L(\sigma)$는

$$
\begin{aligned}
l(\sigma) &\propto \sum_{i=1}^{n} \delta_{(i)} \left[-\ln\sigma - y_{(i)}/\sigma + \ln(1 - (1 - e^{-y_{(i)}/\sigma})^{\beta}) \right] \\
+ & \sum_{i=1}^{n} (1 - \delta_{(i)}) \left[\ln\beta - \ln\sigma + (\beta - 1)\ln(1 - e^{-y_{(i)}/\sigma}) \right]
\end{aligned}
$$

이다. 이로부터 σ에 대한 우도방정식 (likehood equation)을 유도하면 이는 비선형방정식이 되므로 해를 직접 구하기가 힘들다. Kim (2014)에서는 Y^0와 T가 모두 일반화 지수분포를 따를때 σ의 근사해를 구하는 방법을 제안하였다. 따라서 위의 경우는 Kim (2014)의 특별한 경우이다. 이 방법에 의하면 σ의 근사해의 형태는

$$
\hat{\sigma}_2 = \frac{-A + \sqrt{A^2 + 4nB}}{2n} \tag{2.8}
$$

과 같다. 여기서

$$
\begin{aligned}
A &= -\sum_{i=1}^{n} \delta_{(i)} y_{(i)} c_{\beta i} - \sum_{i=1}^{n} (1 - \delta_{(i)}) y_{(i)} (1 - (\beta - 1) a_i) - \sum_{i=1}^{n} y_{(i)} \\
B &= \sum_{i=1}^{n} \delta_{(i)} y_{(i)}^2 d_{\beta i} + \sum_{i=1}^{n} (1 - \delta_{(i)}) y_{(i)}^2 (\beta - 1) b_i
\end{aligned}
$$

이다. $a_j, b_j, c_{\beta j}, d_{\beta j}$, $j = 1, \ldots, n$은

$$
p_j = 1 - \prod_{i \le j} \left(\frac{n - i + 1}{n - i + 2} \right)^{\delta_{(i)}} \left(\frac{n}{n+1} \right)^{1 - \delta_{(1)}}, \quad \xi_j = -\ln(1 - p_j^{1/\beta}), \tag{2.9}
$$

$$
h_1(x_j) = \frac{e^{-x_j}}{1 - e^{-x_j}}, \quad h_2(x_j; \beta) = \frac{e^{-x_j}}{1 - (1 - e^{-x_j})^{\beta}}, \quad x_j = y_{(j)}/\sigma \tag{2.10}
$$

일 때

$$
\begin{aligned}
a_j &= h_1(\xi_j) - h_1'(\xi_j)\xi_j, \quad b_j = -h_1'(\xi_j) \\
c_{\beta j} &= h_2(\xi_j; \beta) - h_2'(\xi_j; \beta)\xi_j, \quad d_{\beta j} = h_2'(\xi_j; \beta)
\end{aligned}
$$

이다. 식 (2.9)의 p_j는 Kaplan-Meier 추정량 (Kaplan-Meier product limit estimator)의 변형된 형태이며 ξ_j는 이를 이용하여 분포의 분위수 (quantile)를 추정한 것이다. Kaplan-Meier 추정량에 대해서는 Kaplan과 Meier (1958), Efron (1967), Meier (1975), Breslow와 Crowley (1974) 등에서 연구되었다. 결국 식 (2.8)의 근사해는 우도방정식에 나타난 식 (2.10)의 함수 h_1, h_2를 추정된 분위수 ξ_j을 중심으로 일차식으로 근사시켜 얻은 결과이다. 이를 이용하면 β가 $\beta \ge 1$일 때는 항상 근사해를 구할 수 있다. 그러나 $\beta < 1$일 때는 자료에 따라 식 (2.8)의 $A^2 + 4nB$가 음이 되어 근사해를 구할 수 없는 경우도 발생한다. 실제의 경우에는 형상모수 β가 대부분 미지이며, σ뿐만 아니라 β에 대해서도 우도함수를 최대로 해야한다 (Kim, 2014). 이 경우 우도함수는

$$
\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} \delta_{(i)} \left(\frac{-(1 - e^{-x_i})^{\beta} \ln(1 - e^{-x_i})}{1 - (1 - e^{-x_i})^{\beta}} \right) + \sum_{i=1}^{n} (1 - \delta_{(i)}) \left(\frac{1}{\beta} + \ln(1 - e^{-x_i}) \right) = 0 \tag{2.11}
$$

이고, 여기서 $x_i = y_{(i)}/\sigma$이다. 위의 식 (2.11)은 Newton-Raphson 방법으로 풀 수 있다. β와 σ의 추정값을 얻기 위해서는 식 (2.8)과 식 (2.11)를 반복적으로 풀어야 한다. 본 논문에서는 σ의 추정량에 관해 살펴보려 하므로 모의실험을 위하여 β를 기지로 가정하였다.

모형 (2.1)에서의 β는 β_1으로 모형 (2.3)에서의 β는 β_2로 두면 식 (2.2), (2.4)에 따라 동일한 중도절단관측의 기대비율 γ에 대해서 $\beta_1 = \gamma/(1-\gamma)$, $\beta_2 = (1-\gamma)/\gamma$가 되므로, 두 값은 역수관계를 갖게된다.

3. 모의실험결과와 자료분석 예제

이 절에서는 모형 (2.1)과 (2.3)이 모수 σ의 추정에 미치는 영향을 모의실험을 통해 살펴보았다. 모수 σ의 실제값은 $\sigma = 0.5, 1, 10$로 하였다. 우선 중도절단관측의 기대비율 γ를 0.5부터 0.05씩 감소하여 0.05까지 고려하였다. 2절에서와 같이 모형 (2.1)에서의 β는 β_1으로 모형 (2.3)에서의 β는 β_2로 두면, β_1과 β_2는 표와 같이 β_1은 1에서 1/19까지 감소하며 β_2는 1에서 19까지 증가한다. 2절에서 언급한 바와 같이 모형 (2.3)을 가정하고 식 (2.8)의 근사해를 이용할 경우 $\beta_2 < 1$일 때, 즉 중도절단관측의 기대비율이 0.5를 넘는 경우에는 근사해를 구하지 못하는 경우도 발생하므로 일단 이 경우는 고려하지 않았다. 표본크기는 $n = 25, 50, 100$으로 하였고 반복수는 $N = 2000$으로 하였다.

Table 3.1, 3.2, 3.3은 각각 $\sigma = 0.5, 1, 10$일 경우이며, 실제모형이 식 (2.1)일 때 식 (2.6), (2.7), (2.8)의 세 추정량을 비교한 것이다. 고려한 모든 σ에 대해서 동일하게 다음과 같은 현상을 관찰할 수 있다. 추정값들은 전반적으로 크게 차이가 나지는 않는다. $\hat{\sigma}_0$과 $\hat{\sigma}_1$은 모든 경우 평균적으로 비슷하나, $\hat{\sigma}_0$의 경우 중도절단관측의 기대비율 γ가 클 때, 즉 완전자료의 비율이 낮을 때 MSE가 상대적으로 큰 경향을 보인다. $\hat{\sigma}_1$과 $\hat{\sigma}_2$에 대해서는 γ가 작을수록, 즉 완전자료의 비율이 높을수록 추정값의 차이가 커지고, 잘못 가정한 모형 그러니까 이 경우는 식 (2.3)을 가정하고 추정했을 때 추정값이 작아지는 경향을 볼 수 있다. 또한 이 경우 전반적으로 추정량의 MSE가 실제모형에서보다 큰 값을 가진다.

Table 3.1 Comparison of estimates when the true model is $1 - F_T = (1 - F_{Y^0})^{\beta_1}$ and the true value is $\sigma = 0.5$

n	γ	β_1	β_2	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	MSE($\hat{\sigma}_0$)	MSE($\hat{\sigma}_1$)	MSE($\hat{\sigma}_2$)
25	0.5	1	1	0.52322	0.49990	0.49990	0.02686	0.00993	0.00993
	0.45	9/11	11/9	0.51573	0.49831	0.51234	0.02299	0.00993	0.01118
	0.4	2/3	3/2	0.51051	0.49768	0.51933	0.01820	0.01028	0.01239
	0.35	7/13	13/7	0.50852	0.49904	0.52492	0.01793	0.01003	0.01337
	0.3	3/7	7/3	0.50441	0.49716	0.52115	0.01662	0.01051	0.01393
	0.25	1/3	3	0.50181	0.49595	0.51809	0.01363	0.00977	0.01376
	0.2	1/4	4	0.50227	0.49790	0.51505	0.01316	0.01024	0.01416
	0.15	3/17	17/3	0.51013	0.50530	0.51640	0.01253	0.00992	0.01407
	0.1	1/9	9	0.50466	0.50184	0.50033	0.01085	0.00953	0.01364
	0.05	1/19	19	0.49838	0.49750	0.48795	0.01076	0.01014	0.01727
50	0.5	1	1	0.51036	0.50063	0.50063	0.01168	0.00511	0.00511
	0.45	9/11	11/9	0.50952	0.50027	0.50784	0.01021	0.00478	0.00512
	0.4	2/3	3/2	0.50411	0.49663	0.50762	0.00871	0.00480	0.00547
	0.35	7/13	13/7	0.50926	0.50422	0.51720	0.00853	0.00506	0.00616
	0.3	3/7	7/3	0.50405	0.50137	0.51300	0.00746	0.00509	0.00663
	0.25	1/3	3	0.50132	0.49814	0.50425	0.00689	0.00498	0.00629
	0.2	1/4	4	0.50572	0.50240	0.50358	0.00655	0.00515	0.00693
	0.15	3/17	17/3	0.50228	0.50016	0.49083	0.00611	0.00510	0.00653
	0.1	1/9	9	0.49816	0.49729	0.47747	0.00554	0.00491	0.00817
	0.05	1/19	19	0.49893	0.49933	0.46701	0.00525	0.00503	0.01064
100	0.5	1	1	0.50528	0.50058	0.50058	0.00518	0.00255	0.00255
	0.45	9/11	11/9	0.50411	0.49997	0.50427	0.00518	0.00264	0.00276
	0.4	2/3	3/2	0.50383	0.49938	0.50456	0.00427	0.00245	0.00269
	0.35	7/13	13/7	0.50308	0.50009	0.50447	0.00404	0.00247	0.00282
	0.3	3/7	7/3	0.50178	0.49986	0.50138	0.00382	0.00250	0.00292
	0.25	1/3	3	0.50107	0.49772	0.49420	0.00335	0.00240	0.00310
	0.2	1/4	4	0.50192	0.50026	0.48907	0.00341	0.00251	0.00320
	0.15	3/17	17/3	0.50072	0.49960	0.47908	0.00289	0.00245	0.00371
	0.1	1/9	9	0.49914	0.49814	0.46504	0.00283	0.00254	0.00589
	0.05	1/19	19	0.49987	0.49959	0.44857	0.00263	0.00250	0.00764

Table 3.2 Comparison of estimates when the true model is $1 - F_T = (1 - F_{Y0})^{\beta_1}$ and the true value is $\sigma = 1$

n	γ	β_1	β_2	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	MSE($\hat{\sigma}_0$)	MSE($\hat{\sigma}_1$)	MSE($\hat{\sigma}_2$)
	0.5	1	1	1.0466	1.0054	1.0054	0.1036	0.0393	0.0393
	0.45	9/11	11/9	1.0346	0.9995	1.0279	0.0895	0.0401	0.0450
	0.4	2/3	3/2	1.0228	0.9932	1.0376	0.0806	0.0395	0.0501
	0.35	7/13	13/7	1.0287	1.0050	1.0523	0.0671	0.0413	0.0516
25	0.3	3/7	7/3	1.0275	1.0072	1.0557	0.0658	0.0409	0.0522
	0.25	1/3	3	1.0074	0.9950	1.0381	0.0553	0.0399	0.0548
	0.2	1/4	4	1.0199	1.0092	1.0467	0.0539	0.0419	0.0648
	0.15	3/17	17/3	1.0034	0.9963	1.0149	0.0462	0.0374	0.0520
	0.1	1/9	9	1.0037	0.9977	0.9964	0.0451	0.0390	0.0560
	0.05	1/19	19	0.9998	0.9989	0.9761	0.0416	0.0396	0.0669
	0.5	1	1	1.0209	0.9974	0.9974	0.0481	0.0194	0.0194
	0.45	9/11	11/9	1.0186	1.0028	1.0182	0.0389	0.0201	0.0218
	0.4	2/3	3/2	1.0132	1.0022	1.0245	0.0358	0.0199	0.0227
	0.35	7/13	13/7	1.0086	0.9983	1.0225	0.0330	0.0197	0.0235
50	0.3	3/7	7/3	1.0054	0.9976	1.0195	0.0305	0.0202	0.0247
	0.25	1/3	3	1.0018	0.9975	1.0106	0.0275	0.0208	0.0260
	0.2	1/4	4	1.0025	0.9955	0.9958	0.0265	0.0208	0.0274
	0.15	3/17	17/3	1.0088	1.0043	0.9906	0.0251	0.0207	0.0292
	0.1	1/9	9	1.0059	1.0032	0.9635	0.0227	0.0199	0.0303
	0.05	1/19	19	1.0018	1.0010	0.9360	0.0211	0.0200	0.0418
	0.5	1	1	1.0147	1.0029	1.0029	0.0212	0.0099	0.0099
	0.45	9/11	11/9	1.0013	0.9962	1.0046	0.0184	0.0098	0.0101
	0.4	2/3	3/2	1.0051	1.0000	1.0102	0.0165	0.0094	0.0102
	0.35	7/13	13/7	1.0083	1.0012	1.0100	0.0168	0.0100	0.0111
100	0.3	3/7	7/3	1.0030	1.0009	1.0047	0.0147	0.0099	0.0119
	0.25	1/3	3	1.0068	1.0016	0.9922	0.0139	0.0101	0.0119
	0.2	1/4	4	1.0064	1.0044	0.9823	0.0135	0.0108	0.0132
	0.15	3/17	17/3	1.0069	1.0036	0.9596	0.0111	0.0095	0.0137
	0.1	1/9	9	1.0038	1.0016	0.9344	0.0110	0.0097	0.0196
	0.05	1/19	19	1.0017	1.0012	0.8949	0.0108	0.0104	0.0279

Table 3.3 Comparison of estimates when the true model is $1 - F_T = (1 - F_{Y0})^{\beta_1}$ and the true value is $\sigma = 10$

n	γ	β_1	β_2	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	MSE($\hat{\sigma}_0$)	MSE($\hat{\sigma}_1$)	MSE($\hat{\sigma}_2$)
	0.5	1	1	10.6226	10.0328	10.0328	12.0999	3.9702	3.9702
	0.45	9/11	11/9	10.4070	10.0331	10.3108	8.7124	3.9702	4.3804
	0.4	2/3	3/2	10.2465	9.9462	10.3733	7.7911	4.0240	4.8592
	0.35	7/13	13/7	10.2058	9.9838	10.4815	6.8622	4.0144	5.3026
25	0.3	3/7	7/3	10.1802	10.0075	10.4760	5.9955	3.8394	4.9719
	0.25	1/3	3	10.2038	10.0636	10.4835	5.9180	3.8356	5.0079
	0.2	1/4	4	10.1524	10.0309	10.3470	5.6397	4.1902	5.6932
	0.15	3/17	17/3	10.1031	10.0053	10.1556	5.1379	4.1006	5.4686
	0.1	1/9	9	10.0722	10.0140	9.9799	4.6654	4.1061	5.7543
	0.05	1/19	19	9.9580	9.9352	9.6838	4.2205	4.0052	6.6265
	0.5	1	1	10.1381	9.9811	9.9811	4.3153	1.9927	1.9927
	0.45	9/11	11/9	10.2676	10.0182	10.1914	4.3391	1.9924	2.1962
	0.4	2/3	3/2	10.1631	10.0150	10.2416	3.8665	2.1730	2.4644
	0.35	7/13	13/7	10.1364	10.0015	10.2488	3.5315	2.0945	2.5050
50	0.3	3/7	7/3	10.0649	9.9942	10.1909	3.0022	1.9961	2.4156
	0.25	1/3	3	10.1404	10.0641	10.2065	2.6501	2.0027	2.5611
	0.2	1/4	4	10.0300	9.9842	9.9970	2.7306	2.1799	3.0163
	0.15	3/17	17/3	10.0911	10.0351	9.8589	2.5931	2.1091	2.6543
	0.1	1/9	9	9.9668	9.9401	9.5242	2.3148	2.0458	3.1572
	0.05	1/19	19	10.0104	9.9996	9.2933	2.0803	1.9455	3.9855
	0.5	1	1	10.1076	9.9812	9.9812	2.1583	0.9898	0.9898
	0.45	9/11	11/9	10.1265	10.0011	10.0809	1.9975	0.9923	1.0381
	0.4	2/3	3/2	10.0845	10.0070	10.1206	1.7853	1.0210	1.0942
	0.35	7/13	13/7	10.0942	10.0511	10.1524	1.5682	0.9769	1.1371
100	0.3	3/7	7/3	10.0108	9.9946	10.0276	1.4816	1.0414	1.2003
	0.25	1/3	3	10.0238	10.0023	9.9224	1.3581	1.0168	1.1943
	0.2	1/4	4	9.9963	9.9777	9.7671	1.3158	1.0379	1.2959
	0.15	3/17	17/3	10.0189	9.9959	9.5744	1.1898	1.0180	1.5389
	0.1	1/9	9	10.0211	10.0048	9.3046	1.1297	1.0236	1.8921
	0.05	1/19	19	9.9982	9.9867	8.8804	1.0812	1.0211	2.9794

Table 3.4, 3.5, 3.6에서 보여주는 바와 같이 실제모형이 식 (2.3)일 때도 추정값의 평균적인 차이는 크지 않다. 그러나 이 경우 모형 (2.1)로 잘못 가정했을 때, 표본크기가 50이상으로 크면 평균적으로 작은 추정값을 보였다. 이러한 추정값의 차이가 중도절단자료의 비율에 따라 크게 달라지지는 않았으나 γ가 감소했을 때 잘못 가정한 모형에서의 추정량의 경우 MSE가 커지는 것을 볼 수 있다. 하지만

Table 3.1-3.3에서 $\hat{\sigma}_1$과 $\hat{\sigma}_2$의 MSE가 차이나는 것보다는 그 정도가 매우 미약하다. 한편 중도절단모형이 모수 σ를 포함하지 않는다고 가정한 $\hat{\sigma}_0$의 경우 Table 3.1-3.3의 경우와 똑같은 경향을 보이고 있으며 MSE가 상대적으로 크게 나타났다.

Table 3.4 Comparison of estimates when the true model is $F_T = (F_{Y0})^{\beta_2}$ and the true value is $\sigma = 0.5$

n	γ	β_1	β_2	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	MSE($\hat{\sigma}_0$)	MSE($\hat{\sigma}_1$)	MSE($\hat{\sigma}_2$)
25	0.5	1	1	0.52886	0.50302	0.50302	0.02742	0.00953	0.00953
	0.45	9/11	11/9	0.51727	0.49942	0.51415	0.02267	0.00848	0.00979
	0.4	2/3	3/2	0.50965	0.49678	0.52021	0.02005	0.00851	0.01103
	0.35	7/13	13/7	0.51061	0.50062	0.52786	0.01828	0.00769	0.01083
	0.3	3/7	7/3	0.51509	0.50091	0.52767	0.01705	0.00703	0.00997
	0.25	1/3	3	0.51424	0.50273	0.52687	0.01599	0.00727	0.00904
	0.2	1/4	4	0.50741	0.50019	0.52153	0.01389	0.00702	0.00776
	0.15	3/17	17/3	0.51119	0.50414	0.52359	0.01280	0.00703	0.00728
	0.1	1/9	9	0.50508	0.49972	0.52054	0.01187	0.00735	0.00690
	0.05	1/19	19	0.49965	0.49727	0.51779	0.01055	0.00807	0.00756
50	0.5	1	1	0.51214	0.50108	0.50108	0.01150	0.00502	0.00502
	0.45	9/11	11/9	0.50868	0.50239	0.51044	0.01065	0.00483	0.00518
	0.4	2/3	3/2	0.50706	0.49906	0.51097	0.00902	0.00410	0.00473
	0.35	7/13	13/7	0.50694	0.49930	0.51292	0.00820	0.00361	0.00440
	0.3	3/7	7/3	0.50881	0.50097	0.51351	0.00783	0.00362	0.00422
	0.25	1/3	3	0.50495	0.49961	0.51096	0.00690	0.00342	0.00372
	0.2	1/4	4	0.50380	0.50015	0.50973	0.00652	0.00350	0.00328
	0.15	3/17	17/3	0.50495	0.50164	0.51143	0.00619	0.00363	0.00335
	0.1	1/9	9	0.50267	0.50035	0.50986	0.00591	0.00390	0.00318
	0.05	1/19	19	0.50081	0.50036	0.51303	0.00555	0.00425	0.00344
100	0.5	1	1	0.50420	0.49807	0.49807	0.00549	0.00258	0.00258
	0.45	9/11	11/9	0.50311	0.49881	0.50320	0.00476	0.00230	0.00242
	0.4	2/3	3/2	0.50289	0.49867	0.50459	0.00416	0.00205	0.00220
	0.35	7/13	13/7	0.50444	0.50002	0.50589	0.00405	0.00202	0.00213
	0.3	3/7	7/3	0.50164	0.49856	0.50444	0.00373	0.00179	0.00182
	0.25	1/3	3	0.50039	0.49905	0.50417	0.00328	0.00167	0.00156
	0.2	1/4	4	0.50069	0.49893	0.50366	0.00313	0.00170	0.00143
	0.15	3/17	17/3	0.50089	0.49994	0.50471	0.00306	0.00184	0.00151
	0.1	1/9	9	0.50263	0.50168	0.50588	0.00285	0.00191	0.00145
	0.05	1/19	19	0.50165	0.50098	0.50632	0.00253	0.00195	0.00148

Table 3.5 Comparison of estimates when the true model is $F_T = (F_{Y0})^{\beta_2}$ and the true value is $\sigma = 1$

n	γ	β_1	β_2	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	MSE($\hat{\sigma}_0$)	MSE($\hat{\sigma}_1$)	MSE($\hat{\sigma}_2$)
25	0.5	1	1	1.0433	1.0414	1.0414	0.0994	0.0516	0.0516
	0.45	9/11	11/9	1.0499	1.0431	1.0139	0.0870	0.0443	0.0419
	0.4	2/3	3/2	1.0410	1.0293	0.9849	0.0784	0.0365	0.0356
	0.35	7/13	13/7	1.0411	1.0326	0.9840	0.0759	0.0375	0.0359
	0.3	3/7	7/3	1.0285	1.0218	0.9708	0.0669	0.0323	0.0314
	0.25	1/3	3	1.0487	1.0395	0.9912	0.0640	0.0342	0.0296
	0.2	1/4	4	1.0357	1.0274	0.9822	0.0584	0.0326	0.0274
	0.15	3/17	17/3	1.0349	1.0294	0.9873	0.0556	0.0345	0.0269
	0.1	1/9	9	1.0356	1.0299	0.9845	0.0520	0.0351	0.0252
	0.05	1/19	19	1.0374	1.0336	0.9877	0.0496	0.0389	0.0268
50	0.5	1	1	1.0202	0.9966	0.9966	0.0468	0.0196	0.0196
	0.45	9/11	11/9	1.0181	0.9983	1.0136	0.0415	0.0178	0.0189
	0.4	2/3	3/2	1.0140	1.0015	1.0259	0.0345	0.0159	0.0184
	0.35	7/13	13/7	1.0152	0.9967	1.0233	0.0340	0.0148	0.0187
	0.3	3/7	7/3	1.0156	1.0036	1.0288	0.0302	0.0141	0.0162
	0.25	1/3	3	1.0071	0.9972	1.0184	0.0263	0.0135	0.0136
	0.2	1/4	4	1.0129	1.0024	1.0212	0.0269	0.0141	0.0132
	0.15	3/17	17/3	1.0058	1.0007	1.0211	0.0246	0.0140	0.0121
	0.1	1/9	9	1.0016	0.9965	1.0155	0.0237	0.0156	0.0124
	0.05	1/19	19	1.0011	0.9985	1.0200	0.0215	0.0164	0.0132
100	0.5	1	1	1.0125	1.0013	1.0013	0.0216	0.0101	0.0101
	0.45	9/11	11/9	1.0096	0.9982	1.0070	0.0194	0.0086	0.0090
	0.4	2/3	3/2	1.0092	1.0011	1.0135	0.0175	0.0084	0.0090
	0.35	7/13	13/7	1.0070	0.9999	1.0130	0.0160	0.0076	0.0080
	0.3	3/7	7/3	1.0065	1.0000	1.0118	0.0146	0.0072	0.0077
	0.25	1/3	3	1.0049	0.9997	1.0093	0.0137	0.0069	0.0065
	0.2	1/4	4	1.0017	0.9985	1.0073	0.0123	0.0068	0.0059
	0.15	3/17	17/3	1.0028	0.9998	1.0080	0.0117	0.0069	0.0056
	0.1	1/9	9	1.0020	0.9997	1.0095	0.0109	0.0071	0.0054
	0.05	1/19	19	0.9997	0.9984	1.0100	0.0105	0.0080	0.0060

Table 3.6 Comparison of estimates when the true model is $F_T = (F_{Y0})^{\beta_2}$ and the true value is $\sigma = 10$

n	γ	β_1	β_2	$\hat\sigma_0$	$\hat\sigma_1$	$\hat\sigma_2$	MSE($\hat\sigma_0$)	MSE($\hat\sigma_1$)	MSE($\hat\sigma_2$)
25	0.5	1	1	10.3522	9.9843	9.9843	9.9427	3.9934	3.9934
	0.45	9/11	11/9	10.4129	9.9850	10.2825	9.0832	3.5924	4.1499
	0.4	2/3	3/2	10.2700	10.0104	10.4584	7.6584	3.2143	4.2429
	0.35	7/13	13/7	10.3190	10.0249	10.5817	7.7082	3.0468	4.3732
	0.3	3/7	7/3	10.2721	10.0181	10.5950	6.6812	2.8527	4.2175
	0.25	1/3	3	10.1798	9.9612	10.4722	5.9698	2.8395	3.6278
	0.2	1/4	4	10.2289	10.0687	10.5197	5.5250	2.8539	3.2453
	0.15	3/17	17/3	10.1782	10.0623	10.4661	5.1763	2.9695	2.9719
	0.1	1/9	9	10.1775	10.0687	10.4674	4.9937	3.1338	2.9675
	0.05	1/19	19	10.0496	9.9888	10.4397	4.3562	3.2546	3.1824
50	0.5	1	1	10.2582	10.0069	10.0069	4.7682	1.9955	1.9955
	0.45	9/11	11/9	10.1796	10.0254	10.1894	4.0132	1.7327	1.9667
	0.4	2/3	3/2	10.1910	10.0351	10.2798	3.7307	1.6868	1.9748
	0.35	7/13	13/7	10.1268	10.0128	10.3017	3.1739	1.3987	1.7521
	0.3	3/7	7/3	10.1221	10.0173	10.2737	3.1407	1.4461	1.6448
	0.25	1/3	3	10.0667	9.9922	10.2121	2.7521	1.3383	1.3339
	0.2	1/4	4	10.1482	10.0441	10.2194	2.5963	1.3895	1.3247
	0.15	3/17	17/3	10.0537	9.9899	10.1784	2.5817	1.5163	1.3128
	0.1	1/9	9	10.0730	10.0207	10.2124	2.3922	1.5441	1.2460
	0.05	1/19	19	10.0382	10.0089	10.2448	2.1997	1.6480	1.3467
100	0.5	1	1	10.1050	10.0188	10.0188	2.0649	0.9900	0.9900
	0.45	9/11	11/9	10.0374	9.9889	10.0756	1.7340	0.8807	0.9087
	0.4	2/3	3/2	10.0889	10.0286	10.1497	1.7866	0.8458	0.9160
	0.35	7/13	13/7	10.0744	9.9757	10.1007	1.6797	0.7606	0.7958
	0.3	3/7	7/3	10.0601	10.0007	10.1052	1.4148	0.6977	0.6920
	0.25	1/3	3	10.0300	9.9798	10.0785	1.3772	0.7080	0.6800
	0.2	1/4	4	10.0376	10.0028	10.0935	1.2619	0.6860	0.5931
	0.15	3/17	17/3	10.0791	10.0339	10.1198	1.2453	0.7265	0.5826
	0.1	1/9	9	10.0226	9.9916	10.0800	1.1198	0.7449	0.5638
	0.05	1/19	19	9.9846	9.9736	10.1085	1.0897	0.8252	0.5963

두 경우 모두에서 MSE는 대략 표본크기에 반비례로 줄어들고 σ^2에 비례해서 커지는 것을 볼 수 있다. 또한 γ가 작을 때 잘못 가정한 모형의 MSE가 상대적으로 큰 경향을 보인다. 이는 γ가 작을수록 실제모형과 가정한 모형이 크게 차이가 나기 때문이 아닌가 생각된다. 실제로 $\gamma = 0.5$일 때는 두 모형 (2.1)과 (2.3)는 동일하다.

2절의 결과를 실제 자료에 적용하여 보자. 다음 자료는 급성골수백혈병 (acute myelogenous leukemia)을 치료하기 위한 보전화학요법 (maintenance chemotheraph)의 효능을 알아보기 위하여 실행한 임상실험결과 중 일부이다. 이 연구는 Embury 등 (1977)에서 행해졌고, Table과 Kim (2004)에서도 분석하였다. 환자를 두 그룹으로 나누어 한 그룹에는 보전화학요법을 행하고, 다른 그룹에는 행하지 않았다. 아래는 보전화학요법 치료를 받은 그룹의 완화시간을 나타내는 자료이며, 주단위로 나타낸 것이다. +는 중도절단을 의미한다.

$$9,\ 13,\ 13+,\ 18,\ 23,\ 28+,\ 31,\ 34,\ 45+,\ 48,\ 161+$$

지수분포를 가정하고, 중도절단분포에 식 (2.1)을 가정하면, $\hat\beta_1 = 4/7 = 0.571$로 추정하여 $\hat\sigma_1$은 식 (2.6)과 동일하게 되며, 이 경우 $\hat\sigma_1 = 60.429$, $\hat\lambda_1 = 1/\hat\sigma_1 = 0.0165$가 된다. 만일 중도절단분포에 식 (2.3)을 가정하고, 식 (2.8)과 (2.11)로부터 β_2와 σ를 추정하면 $\hat\beta_2 = 1.748$, $\hat\sigma_2 = 64.322$, $\hat\lambda_2 = 1/\hat\sigma_2 = 0.0155$가 된다. 추정된 $\hat\beta_2$은 $1/\hat\beta_1 = 7/4 = 0.175$와 거의 일치하는 것을 볼 수 있다. 그러나 이는 일반적인 것은 아니고 이 자료의 경우 나타난 결과로 보인다.

4. 결론 및 토의

본 논문에서는 임의중도절단 자료의 분석에 중도절단모형이 어떤 영향을 미치는지에 대해 생각해 보고자 하였다. 실제 자료분석시에 중도절단모형을 정확하게 정하기는 쉽지 않으므로 이러한 연구가 의미있다고 생각한다.

자료에 지수분포를 가정하고, 중도절단모형으로는 잘 알려진 Koziol-Green 모형과 일반화 지수분포 모형을 고려하였다. 이 경우 지수분포의 척도모수 추정에 이 모형이 어떤 영향을 주는지에 대해서 살펴보았다. Koziol-Green 모형의 경우, 지수분포의 척도모수 추정이 상대적으로 쉽지만, 일반화 지수분포 모형의 경우에는 최대우도추정량이 해석적으로 구해지지 않으므로 근사적으로 구해야 한다.

모의실험을 통해 관찰한 결과 추정값이 평균적으로는 중도절단모형에 따라 크게 좌우되지 않는 것으로 보인다. 그러나 중도절단자료의 기대비율이 작을수록 위의 두 모형은 점점 더 차이가 나게 되며 가정한 모형과 실제모형이 심하게 다르게 된다. 이 경우 잘못된 모형을 가정했을 때 추정량의 MSE가 상대적으로 커지는 것을 볼 수 있었다.

본 연구는 모의실험을 통하여 중도절단모형이 추정에 주는 영향을 살펴본 것으로 차후에 좀 더 이론적인 접근을 통한 연구가 필요하다. 또한 지수분포 외에 자주 이용되는 다른 일반적인 분포로의 확장이나 일반화도 필요하다고 생각한다.

References

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorships. *The Annals of Statistics*, **2**, 437-453.

Chen, Y. Y., Hollander, M. and Langberg, N. A. (1982). Small-sample results for the Kaplan Meier estimator. *Journal of the American Statistical Association*, **77**, 141-144.

Cho, Y., Lee, C. and Shin, H. (2013). Estimation for the generalized exponential distribution under progressive type I interval censoring. *Journal of the Korean Data & Information Science Society*, **24**, 1309-1317.

Cohen, A. C. (1991). *Truncated and censored samples, theory and applications*, Marcel Dekker, Inc., New York.

Csörgő, S. and Horváth, L. (1981). On the Koziol-Green model for random censorship. *Biometrika*, **68**, 391-401.

Efron, B. (1967). The two sample problem with censored data. *Proceedings of the 5th Berkeley Symposium*, **4**, 831-853.

Embury, S. H., Elias, L., Heller, P. H., Hood, C. E., Greenberg, P. L. and Schrier, S. L. (1977). Remission maintenance therapy in acute myelogenous leukemia. *Western Journal of Medicine*, **126**, 267-272.

Gupta, R. D. and Kundu, D. (1999). Generalized exponential distributions. *Australian and New Zealand Journal of Statistics,* **41**, 173-188.

Gupta, R. D. and Kundu, D. (2001a). Exponentiated exponential family: An alternative to gamma and Weibull. *Biometrical Journal*, **43**, 117-130.

Gupta, R. D. and Kundu, D. (2001b). Generalized exponential distributions: Different methods of estimation. *Journal of Statistical Computation and Simulation*, **69**, 315-338.

Gupta, R. D. and Kundu, D. (2002). Generalized exponential distribution: Statistical inferences. *Journal of Statistical Theory and Applications*, **1**, 101-118.

Gupta, R. D. and Kundu, D. (2007). Generalized exponential distribution: Existing results and some recent developments. *Journal of Statistical Planning and Inference*, **137**, 3537-3547.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.

Kim, N. (2014). Approximate MLE for the scale parameter of the generalized exponential distribution under random censoring. *Journal of The Korean Statistical Society*, **43**, 119-131.

Koziol, J. A. and Green, S. B. (1976). A Cramér-von Mises statistic for randomly censored data. *Biometrika*, **63**, 465-474.

Lee, E. T. and Wang, J. W. (2003). *Statistical methods for survival data analysis*, John Wiley & Sons, Inc., New Jersey.

Meier, P. (1975). Estimation of a distribution function from incomplete observations. In *Perspectives in Probability and Statistics*, edited by J. Gani, Academic Press, London, 67-87.

Raqab, M. Z. (2002). Inference for generalized exponential distribution based on record statistics. *Journal of Statistical Planning and Inferences*, **104**, 339-350.

Tableman, M. and Kim, J. S. (2004). *Survival analysis using S*, Champman & Hall/CRC, Boca Raton.

Journal of the Korean Data & Information Science Society
2014, **25**(2), 393–402

http://dx.doi.org/10.7465/jkdi.2014.25.2.393
한국데이터정보과학회지

The influence of the random censorship model on the estimation of the scale parameter of the exponential distribution†

Namhyun Kim[1]

[1]Department of Science, Hongik University

Received 28 January 2014, revised 20 February 2014, accepted 11 March 2014

Abstract

The simplest and the most important distribution in survival analysis is the exponential distribution. In this paper, we investigate the influence of the random censorship model on the estimation of the scale parameter of the exponential distribution. The considered random censorship models are Koziol-Green model and the generalized exponential distribution model. Two models have different meanings. Through the simulation study, the averages of the estimated values of the parameter do not show big differences, however the MSE of the estimator tends to be bigger when the supposed model is significantly different from the true model.

Keywords: Koziol-Green model, maximum likelihood estimator, random censorship, random censorship model.

---

† This work was supported by 2013 Hongik University Research Fund.
[1] Professor, Department of Science, Hongik University, Seoul 121-791, Korea.
E-mail: nhkim@hongik.ac.kr