

## 정보적 중도절단을 고려한 최대 편우도 추정량의 정규성

정대현\*, 원동유\*\*

\*충북 청주시 흥덕구 개신동 산 48 충북대학교 자연과학대학 통계학과 부교수

\*\*충북 청주시 흥덕구 개신동 산 48 충북대학교 자연과학대학 전자계산학과 박사

## Normality of the MPLE of a Proportional Hazard Model for Informative Censored Data

Daehyun Chung\*, DongYu Won\*\*

\*Dept, of Statistics, Chungbuk National University, Cheongju, Chungbuk

\*\*Dept, of Computer Science, Chungbuk National University, Cheongju, Chungbuk

### Abstract

We study the normality of the maximum partial likelihood estimators for the proportional hazard model with informative censored data. The proposed models cover the cases in which the times to a primary event may be informatively or randomly censored and the times to a secondary event may be randomly censored.

To estimate the parameters and to check the normality of the parameters in the model, we adopt the partial likelihood and counting process to use the martingale central limit theorem. Simulation studies are performed to examine the normality of the MPLE's for the five cases in which they depend upon the proportions of randomly censored and informative censored data.

Key Words : proportional hazard model, random censored, informative censored, martingales, counting process.

## 1. 서론

생존분석에서 생존시간을 관측하고자할 때 오른쪽 중도절단 자료는 흔히 나타나는 자료 중의 하나이다. 대표적인 중도절단의 형태는 정보적 중도절단과 임의 중도절단이 대부분 연구에서는 정보적 중도절단을 임의의 중도절단으로 처리하고 있다. 실패사건과 중도절단자료의 독립성을 가정하는 것이 무리일 경우 중도절단 자료를 무시한다면 과대 또는 과소 생존함수의 추정량이 얻어진다. (Lagakos (1979)).

일반적으로 생존함수의 추론에서는 정보적 중도절단을 고려하기 위하여 실패사건과 정보적 중도절단의 결합분포를 가정하지만 정보적 중도절단을 실패사건에 정보를 제공하는 또는 실패사건의 형태로 간주하여 경쟁위험모형을 적용하여 분석하고자 한다.

정보적 중도절단 자료를 고려한 모수적 방법으로는 Marshall과 Olkin(1967)이  $T_1$ (실패사건 시간)과  $T_2$ (중도절단 시간)이 종속적이며  $(T_1, T_2)$ 가 이변량 지수분포를 따를 때 생존함수를 추정하였다. Nádas(1971)는 이변량 정규분포에서 조생존함수가 순생존함수를 결정한다는 것을 증명하였고 다변량 정규분포임을 가정한 경우 모수의 추정에 관한 문제는 Smith와 Helms(1995)에 의하여 다루어졌다.

대표적인 준모수적 접근 방법을 제시한 Klein, Keiding과 Kamby(1989)는 서로 다른 두 지점에서 발생하는 두 종류의 실패사건을 모형화하기 위하여 Marshall과 Olkin (1967)의 준모수 모형으로부터 주변위험 발생에 초점을 두어 비례위험 모형을 제안하였다.

Klein, Keiding과 Kamby의 모형을 확장하여 두 종류의 실패사건 사이에 비례위험모형을 새로이 제안하여 MPLE를 이용하여 추론한 추정량의 정규성을 연구하고자 한다.

두 종류의 실패사건의 구체적인 예를 백혈병 환자가 골수 이식 수술을 받는 것을 유형 1의 실패사건, 백혈병 환자가 재발되거나 사망하는 사건을 유형 2의 실패사건으로 보는 경우이다. 첫 번째, 백혈병이 재발되어 새로운 방법으로 치료를 받거나 사망한 경우에는 정보적 중도절단된 자료를 얻게 된다. 골수 이식을 받기 전에 사망하거나 재발하지 않은 환자는 유형1의 실패사건인 골수 이식 수술을 받게 되고 차후 유형 2의 실패사건인 사망하거나 재발을 경험하게 된다. 골수이식 수술을 받기 전에 재발하지도 않고 사망하지도 않은 환자는 임의적 중도절단이 될 수도 있다. 마지막으로 골수이식을 받은 후에 재발 또는 사망하기도 전에 임의적 중도절단이 발생할 수도 있다.

따라서 골수 이식을 받을 때까지 기다리는 시간이 환자의 사망 또는 재발에 어떠한 영향을 미치는지 알아보는 것은 매우 중요하며, MPLE의 정규성의 변화를 알아보고 상황에 따른 MPLE의 효율성을 연구하여 실제 자료의 분석에 응용할 수 있다.

## 2. 정보적 중도절단을 고려한 비례위험 모형

이 절에서는 경쟁위험모형에서 Kalbfleisch와 Prentice(1980)의 방법을 이용하여 정보적 중도절단과 비례위험모형을 고려해 보기로 하자. 각각의 관측대상이 상호 배반인 두 가지 유형의 사건 중 하나 또는 모두를 경험한다고 가정하자.  $T_v$  ( $v=1,2$ )를 관측대상이 사건  $v$ 를 경험하는 시간이라면  $X = \min(T_1, T_2)$ 는 첫 번째 사건을 경험하는 시간이다.  $(T_1, T_2)$

중도 절단된 환자이며 골수 이식을 받은 후에 재발하거나 사망한 환자는 범주 4에 속하게 된다. 그리고, 만일

$i$ 번째 환자의 실패사건 시간을  $T_i$ , 정보적 중도절단 시간을  $U_i$ , 임의 중도절단 시간을  $C_i$ , 공변량 벡터를  $Z_i$ 라고 하자. 그리고

$$\begin{aligned} X_{1i} &= \min\{T_i, U_i, C_i\}, \\ X_{2i} &= \min\{U_i, C_i\}, \\ \delta_i &= I\{T_i \leq U_i \wedge C_i\}, \\ \xi_i &= I\{U_i \wedge C_i\}. \end{aligned}$$

계수과정  $N_{1i}(t)$ ,  $N_{2i}(t)$ ,  $Y_{1i}(t)$ ,  $N_{2i}(t)$  그리고  $Y_{2i}(t)$ 를 다음과 같이 정의하자.

$$\begin{aligned} N_{1i}(t) &= I\{X_{1i} \leq t, \delta_i = 1\}, \\ N_{2i}(t) &= I\{X_{1i} \leq t, \delta_i = 0, \xi_i = 1\}, \\ Y_{1i}(t) &= I\{X_{1i} \geq t\}, \\ N_{2i}(t) &= I\{X_{2i} \leq t, \delta_i = 1, \xi_i = 1\}, \\ Y_{2i}(t) &= I\{X_{1i} < t, X_{2i} \geq t\}. \end{aligned}$$

주어진 계수과정  $N(\cdot)$ 에 대하여,  $N(u) - N(u-) = 1$ 이면  $\Delta N(u) = 1$ 이라고 하고, 그 외에는  $\Delta N(u) = 0$ 이라고 정의하자.

Fleming과 Harrington(1991)을 참조하여 모형 식 (2.2)을 가정하여  $n$ 명의 환자로부터 구성할 수 있는 편우도 함수는 다음과 같이 표현된다.

$$\begin{aligned} PL_n(\theta) &= \prod_{s=1}^2 \prod_{i=1}^n \prod_{0 \leq u} \left[ \frac{Y_{1i}(u) \exp(\mathbf{z}_i \boldsymbol{\beta})}{\sum_{j=1}^n Y_{1j}(u) \exp(\mathbf{z}_j \boldsymbol{\beta})} \right]^{\Delta N_{1i}(u)} \\ &\quad \times \prod_{i=1}^n \prod_{0 \leq u} \left[ \frac{Y_{2i}(u) \exp(\mathbf{z}_i \boldsymbol{\beta} + x_{1i} \alpha)}{\sum_{j=1}^n Y_{2j}(u) \exp(\mathbf{z}_j \boldsymbol{\beta} + x_{1j} \alpha)} \right]^{\Delta N_{2i}(u)} \end{aligned}$$

일반적으로  $\theta$ 의 최대 편우도 추정량은  $U(\theta) = \frac{\partial \ln PL_n(\theta)}{\partial \theta} = 0$  의 해가 되므로

$\theta = (\boldsymbol{\beta}, \alpha)$ 의 추정량을 구하기 위하여 위 식을 이용한다.

로그 편우도 함수가 오복함수임으로  $\theta$ 의 최대 편우도 추정량의 유일성을 보장된다.

#### 4. 최대 편우도 추정량의 근사 분포

관측구간을  $(0, t]$ 로 제한한 경우

$$U(\theta, t) = U(\beta, \alpha, t) = \frac{\partial \ln PL_n(\beta, \alpha, t)}{\partial(\beta, \alpha)} \text{ 라고 하자.}$$

$$\begin{aligned} \text{단, } PL_n(\beta, \alpha, t) &= \prod_{s=1}^2 \prod_{i=1}^n \prod_{0 \leq u \leq t} \left[ \frac{Y_{1i}(u) \exp(z_i \beta)}{\sum_{j=1}^n Y_{1j}(u) \exp(z_j \beta)} \right]^{\Delta N_{1i}(u)} \\ &\times \prod_{i=1}^n \prod_{0 \leq u \leq t} \left[ \frac{Y_{2i}(u) \exp(z_i \beta + x_{1i} \alpha)}{\sum_{j=1}^n Y_{2j}(u) \exp(z_j \beta + x_{1j} \alpha)} \right]^{\Delta N_{2i}(u)} \end{aligned}$$

모수 벡터  $\theta = (\beta, \alpha)$ 의 최대 편우도 추정량의 근사 분포는  $\theta_0$ 는  $\theta$ 의 참값이며  $\hat{\theta}$ 은  $\theta$ 의 최대 편우도 추정량이라고 할 때 마팅게일 중심극한정리를 사용하여  $n^{-1/2} U(\theta_0, t)$ 가 Gaussian 과정으로 수렴함을 보이고,  $U(\hat{\theta}, t)$ 의 Taylor 급수 전개를 이용하여 정규성을 증명하고자 한다.

##### 정리 4.1

$N_{11i}, N_{12i}, N_{2i}, Y_{1i}, Y_{2i}, Z_i, X_{1i}, \beta, \alpha$ 는 3장에서 정의된 것과 같이 정의하고,

$$\begin{aligned} H_{1i}(u) &= z_i - \frac{\sum_{j=1}^n Y_{1j}(u) z_j e^{z_j \beta_0}}{\sum_{j=1}^n Y_{1j}(u) e^{z_j \beta_0}}, \\ H_{21i}(u) &= z_i - \frac{\sum_{j=1}^n Y_{2j}(u) z_j e^{z_j \beta_0 + x_{1j} \alpha_0}}{\sum_{j=1}^n Y_{2j}(u) e^{z_j \beta_0 + x_{1j} \alpha_0}}, \quad H_{22i}(u) = x_{1i} - \frac{\sum_{j=1}^n Y_{2j}(u) x_{1j} e^{z_j \beta_0 + x_{1j} \alpha_0}}{\sum_{j=1}^n Y_{2j}(u) e^{z_j \beta_0 + x_{1j} \alpha_0}} \end{aligned}$$

라고 하자.

위  $H(u)$ 들이 마팅게일 중심극한정리의 가정을 만족한다고 가정하면,  $n \rightarrow \infty$ 일때  $n^{-1/2} U(\theta_0, t)$ 은 평균이 0인 이변량 정규분포로 수렴한다.

##### 증명

$U(\theta_0, t)$ 를 다음과 같이 표현 할 수 있다.

$$U(\theta_0, t) = \frac{\partial \phi_1(t)}{\partial \theta_0} + \frac{\partial \phi_2(t)}{\partial \theta_0} + \frac{\partial \phi_3(t)}{\partial \theta_0} = \frac{\partial \phi(t)}{\partial \theta_0}.$$

단,  $\phi_i(t)$ 는 관측값의 범위가 구간  $(0, t]$ 에 제한된  $\phi_i$ 이다. 그러면, 다음과 같이 나타낼 수 있다.

$$U(\theta_0, t) = \begin{pmatrix} \frac{\partial \phi(t)}{\partial \beta_0} \\ \frac{\partial \phi(t)}{\partial \alpha_0} \end{pmatrix} = \begin{pmatrix} U_b \\ U_a \end{pmatrix}.$$

$$\text{단, } U_b = \sum_{i=1}^n \int_0^t \left\{ z_i - \frac{\sum_{j=1}^n Y_{1j}(u) z_j e^{z_j \beta_0}}{\sum_{j=1}^n Y_{1j}(u) e^{z_j \beta_0}} \right\} d[N_{11i}(u) + N_{12i}(u)]$$

$$+ \sum_{i=1}^n \int_0^t \left\{ z_i - \frac{\sum_{j=1}^n Y_{2j}(u) z_j e^{z_j \beta_0 + x_{1j} \alpha_0}}{\sum_{j=1}^n Y_{2j}(u) e^{z_j \beta_0 + x_{1j} \alpha_0}} \right\} dN_{2i}(u) ,$$

$$U_a = \sum_{i=1}^n \int_0^t \left\{ x_{1i} - \frac{\sum_{j=1}^n Y_{2j}(u) x_{1j} e^{z_j \beta_0 + x_{1j} \alpha_0}}{\sum_{j=1}^n Y_{2j}(u) e^{z_j \beta_0 + x_{1j} \alpha_0}} \right\} dN_{2i}(u) ,$$

$$E_1(\beta_0, u) = \frac{\sum_{j=1}^n Y_{1j}(u) z_j e^{z_j \beta_0}}{\sum_{j=1}^n Y_{1j}(u) e^{z_j \beta_0}} ,$$

$$E_{21}(\beta_0, \alpha_0, u) = \frac{\sum_{j=1}^n Y_{2j}(u) z_j e^{z_j \beta_0 + x_{1j} \alpha_0}}{\sum_{j=1}^n Y_{2j}(u) e^{z_j \beta_0 + x_{1j} \alpha_0}} ,$$

$$E_{22}(\beta_0, \alpha_0, u) = \frac{\sum_{j=1}^n Y_{2j}(u) x_{1j} e^{z_j \beta_0 + x_{1j} \alpha_0}}{\sum_{j=1}^n Y_{2j}(u) e^{z_j \beta_0 + x_{1j} \alpha_0}} ,$$

$$\begin{aligned} M_{1i}(t) &= N_{11i}(t) + N_{12i}(t) - \int_0^t Y_{1i}(u) e^{z_i \beta_0} (\lambda_{011}(u) + \lambda_{012}(u)) du \\ &= N_{1i}(t) - A_{1i}(t) , \end{aligned}$$

$$\begin{aligned} M_{2i}(t) &= N_{2i}(t) - \int_0^t Y_{2i}(u) e^{z_i \beta_0 + x_{1i} \alpha_0} \lambda_{02}(u) du \\ &= N_{2i}(t) - A_{2i}(t) . \end{aligned}$$

Fleming과 Harrington(1991)의 1.3절에서 제안한  $M_{1i}(t)$ 와  $M_{2i}(t)$ 는 마팅게일이고,

$$\frac{\sum_{j=1}^n (z_j - E_1(\beta_0, u)) Y_{1j}(u) z_j e^{z_j \beta_0}}{\sum_{j=1}^n Y_{1j}(u) e^{z_j \beta_0}} = 0 ,$$

$$\frac{\sum_{j=1}^n (z_j - E_{21}(\beta_0, \alpha_0, u)) Y_{2j}(u) z_j e^{z_j \beta_0 + x_{1j} \alpha_0}}{\sum_{j=1}^n Y_{2j}(u) e^{z_j \beta_0 + x_{1j} \alpha_0}} = 0,$$

$$\frac{\sum_{j=1}^n (z_j - E_{22}(\beta_0, \alpha_0, u)) Y_{2j}(u) x_{1j} e^{z_j \beta_0 + x_{1j} \alpha_0}}{\sum_{j=1}^n Y_{2j}(u) e^{z_j \beta_0 + x_{1j} \alpha_0}} = 0$$

이므로,  $U_b$ 와  $U_a$ 는 다음과 같은 마팅게일 구조를 가지고 있다.

$$U_b = \sum_{i=1}^n \int_0^t H_{1i}(u) dM_{1i}(u) + \sum_{i=1}^n \int_0^t H_{21i}(u) dM_{2i}(u),$$

$$U_a = \sum_{i=1}^n \int_0^t H_{22i}(u) dM_{2i}(u).$$

간단하게 표현하기 위해  $n^{-1/2} U_b = y_1 + y_2$ 와  $n^{-1/2} U_a = y_3$ 로 놓자.

이 때  $y_1, y_2, y_3$ 는 다음과 같다.

$$y_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t H_{1i}(u) dM_{1i}(u),$$

$$y_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t H_{21i}(u) dM_{2i}(u),$$

$$y_3 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t H_{22i}(u) dM_{2i}(u).$$

$n^{-1/2} \mathbf{U}(\boldsymbol{\theta}_0, t) = n^{-1/2}(U_b, U_a) = (y_1 + y_2, y_3)$ 의 근사적인 이변량 정규성은

사실 1. 임의의 상수  $a_1$ 과  $a_2$ 에 대하여,  $(y_1, a_1 y_2 + a_2 y_3)$ 가 마팅게일 중심극한정리를 이용하여 평균이 0인 근사적 이변량 정규분포를 따른다.

사실 2.  $(y_1, y_2, y_3)$ 는 평균이 0인 근사적 삼변량 정규분포를 따른다.

따라서,  $n^{-1/2}(U_b, U_a) = (y_1 + y_2, y_3) \sim N\left(0, \begin{bmatrix} \sigma_{y_1}^2 + \sigma_{y_2}^2 & 0 \\ 0 & \sigma_{y_3}^2 \end{bmatrix}\right)$ 이며 여기서,

$$\begin{aligned}
 \sigma_{y_1}^2 &= n^{-1} E \left\{ \sum_{i=1}^n \int_0^t [H_{1i}(u)]^2 dA_{1i}(u) \right\} \\
 &= n^{-1} E \left\{ \sum_{i=1}^n \int_0^t \left( z_i^2 - \frac{2z_i \sum_{j=1}^n Y_{1j}(u) z_j e^{z_j \beta_0}}{\sum_{j=1}^n Y_{1j}(u) e^{z_j \beta_0}} - \left( \frac{\sum_{j=1}^n Y_{1j}(u) z_j e^{z_j \beta_0}}{\sum_{j=1}^n Y_{1j}(u) e^{z_j \beta_0}} \right)^2 \right) \right. \\
 &\quad \left. \times Y_{1i}(u) e^{z_i \beta_0} \lambda_{01}(u) du \right\} \\
 &= n^{-1} E \left\{ \int_0^t \left( \sum_{i=1}^n Y_{1i}(u) z_i^2 e^{z_i \beta_0} - \frac{\left( \sum_{j=1}^n Y_{1j}(u) z_j e^{z_j \beta_0} \right)^2}{\sum_{j=1}^n Y_{1j}(u) e^{z_j \beta_0}} \right) \lambda_{01}(u) du \right\} \\
 \sigma_{y_2}^2 &= \frac{1}{n} E \left\{ \sum_{i=1}^n \int_0^t [H_{21i}(u)]^2 dA_{2i}(u) \right\} \\
 &= \frac{1}{n} E \left\{ \int_0^t \left( \sum_{i=1}^n Y_{2i}(u) z_i^2 e^{z_i \beta_0 + x_{1i} \alpha_0} - \frac{\left( \sum_{j=1}^n Y_{2j}(u) z_j e^{z_j \beta_0 + x_{1j} \alpha_0} \right)^2}{\sum_{j=1}^n Y_{2j}(u) e^{z_j \beta_0 + x_{1j} \alpha_0}} \right) \lambda_{02}(u) du \right\} \\
 \sigma_{y_3}^2 &= \frac{1}{n} E \left\{ \sum_{i=1}^n \int_0^t [H_{22i}(u)]^2 dA_{2i}(u) \right\} \\
 &= \frac{1}{n} E \left\{ \int_0^t \left( \sum_{i=1}^n Y_{2i}(u) x_{1i}^2 e^{z_i \beta_0 + x_{1i} \alpha_0} - \frac{\left( \sum_{j=1}^n Y_{2j}(u) x_{1j} e^{z_j \beta_0 + x_{1j} \alpha_0} \right)^2}{\sum_{j=1}^n Y_{2j}(u) e^{z_j \beta_0 + x_{1j} \alpha_0}} \right) \lambda_{02}(u) du \right\}
 \end{aligned}$$



마지막으로, 최대 편우도 추정량( $\hat{\theta}$ )의 근사적 정규성에 대하여 알아보자.

**정리 4.2**

$\hat{\theta}$ 을  $\theta$ 의 최대 편우도 추정량이라고 하고  $\theta_0$ 을  $\theta$ 의 참값이라고 하자.

$$\Sigma(\theta_0, t) = \begin{bmatrix} \sigma_{y_1}^2 + \sigma_{y_2}^2 & 0 \\ 0 & \sigma_{y_3}^2 \end{bmatrix} \text{이라고 하자.}$$

정리 4.1의 조건을 가정하면,

(1)  $\Sigma(\theta_0, t)$ 는 정칙(nonsingular)이다.  $\Sigma(\theta_0, t)$ 의 일치 추정량은  $n^{-1} I(\hat{\theta}, t)$ 에 의하여 제공되었다. 단,  $I(\theta, t)$ 는 로그 편우도함수의 이차 편도함수에 (-1)을 곱한 것으로 정의된 정보 행렬이다.

(2)  $\hat{\theta}$ 는  $\theta_0$ 의 일치추정량이다. 즉,  $\hat{\theta} \xrightarrow{p} \theta_0$ .

(3)  $n \rightarrow \infty$  일 때  $n^{1/2}(\hat{\theta} - \theta_0)$ 는 평균이 0이고 공분산 행렬이  $\Sigma^{-1}(\theta_0, t)$ 인 이변량 정규분포로 수렴한다.

증명

$n^{-1}I(\hat{\theta}, t)$ 와  $\hat{\theta}$ 의 일치성은 Fleming과 Harrington(1991)의 정리 8.2.1과 정리 8.3.1의 증명과 유사한 방법으로 보일 수 있다.  $\hat{\theta}$ 의 근사적 정규성은 정리 4.1과 Fleming과 Harrington(1991)의 정리 8.3.2의 증명과 유사한 방법으로 보일 수 있다.  $U(\hat{\theta}, t)$ 의 Taylor 급수 전개에 첫 번째를 사용하면

$$\begin{aligned} U(\hat{\theta}, t) &= U(\theta_0, t) + \frac{\partial}{\partial \theta} U(\theta, t) \Big|_{\theta = \theta^*} (\hat{\theta} - \theta_0) \\ &= U(\theta_0, t) - I(\theta^*, t) (\hat{\theta} - \theta_0) \end{aligned}$$

와 같이 나타낼 수 있다. 단,  $\theta^*$ 는  $\hat{\theta}$ 과  $\theta_0$  사이의 선상에 있다.  $U(\hat{\theta}, t) = 0$  이므로 위의 식은 다음과 같다.

$$n^{-1}I(\theta^*, t) n^{1/2}(\hat{\theta} - \theta_0) = n^{-1/2}U(\theta_0, t).$$

$n^{-1/2}U(\theta_0, t)$ 는 정리 3.3에서 근사적으로 공분산 행렬이  $\Sigma(\theta_0, t)$ 인 정규분포를 따름을 보였다.  $\hat{\theta}$ 는  $\theta_0$ 의 일치추정량이므로  $n \rightarrow \infty$ 일 때  $\theta^* \xrightarrow{p} \theta_0$  이고, Fleming 과 Harrington(1991)의 정리 8.2.1의 (2)에서  $n \rightarrow \infty$ 일 때  $I(\theta^*, t) \xrightarrow{p} \Sigma(\theta_0, t)$ 함을 보였으므로  $n \rightarrow \infty$ 일 때  $n^{1/2}(\hat{\theta} - \theta_0)$ 는 평균이 0이고 공분산 행렬이  $\Sigma^{-1}(\theta_0, t)$ 인 이변량 정규분포로 수렴함을 증명하였다. ■

## 5. 최대 편우도 추정량의 정규성에 대한 모의실험 및 결론

### 5.1 모의실험의 절차

이 장에서는 3장에서 설명한 4가지 범주에 대하여 모의실험을 통하여 최대 편우도 추정량의 정규성을 다음 5가지 경우에 대하여 조사하였다.

- 경우 1. 임의 중도절단 자료의 비율이 가장 큰 경우,
- 경우 2. 정보를 가진 중도절단된 자료의 비율이 가장 큰 경우,
- 경우 3. 첫 번째 사건이 관측된 후 임의 중도절단 자료의 비율이 가장 큰 경우,
- 경우 4. 첫 번째 사건이 관측된 후 두 번째 사건이 관측된 자료의 비율이 가장 큰 경우,
- 경우 5. 위 4가지 경우의 비율이 비슷한 경우.

각 경우 최대 편우도 추정량의 영향을 알아보기 위하여  $\alpha$ 와  $\beta$ 의 최대 편우도 추정량에 대한 히스토그램을 이용하여 최대 편우도 추정량의 표본분포들의 근사적 정규성의 적합성에 대하여 조사하였다.



앞에서 정의된 공변량  $Z = z$ 가 주어졌을 때 3변량  $(T, U, C)$ 의 조건부 분포로부터 표본  $n(=50, 100)$ 개를 추출하여 모의 실험하는 과정을  $N(=1000)$ 번 반복 실행하였다.  $Z$ 의 값  $z$ 는  $p=0.5$ 인 베르누이 분포로부터 얻었다.  $\lambda_{01}, \lambda_{02}, \beta, \alpha$ 가 주어졌을 때  $T=x_1, U=u_1, U=u_2 | T=x_1$ 에 대한 조건부 위험률은 각각 다음과 같이 가정했다.

$$\lambda_T(x_1 | z) = \lambda_{01} e^{z\beta},$$

$$\lambda_U(u_1 | z) = \lambda_{02} e^{z\beta},$$

$$\lambda_{U|T}(u_2 | x_1, z) = \lambda_{02} e^{z\beta + x_1\alpha} I(u_2 > x_1).$$

$T$ 와  $U$ 의 결합밀도함수는 아래와 같이 나타낼 수 있다.

$$\begin{aligned} f_{T,U}(x_1, u_2 | z) &= \lambda_{01}\lambda_{02} e^{x_1\alpha + 2z\beta} e^{-(\lambda_{01} + \lambda_{02})e^{z\beta}x_1 - \lambda_{02}e^{z\beta + x_1\alpha}(u_2 - x_1)} I(u_2 > x_1) \\ &= f_T(x_1 | z) f_{U|T}(u_2 | x_1, z). \end{aligned}$$

$$\begin{aligned} \text{단, } f_T(x_1 | z) &= P(T = x_1, U > x_1) & \pi &= \frac{\lambda_{01}}{\lambda_{01} + \lambda_{02}}, \\ &= \pi \lambda e^{\lambda x_1}, & \lambda &= (\lambda_{01} + \lambda_{02}) e^{z\beta}, \end{aligned}$$

$$\begin{aligned} f_{U|T}(u_2 | x_1, z) &= \lambda_{02} e^{x_1\alpha + z\beta} \exp\left\{-\int_{x_1}^{u_2} \lambda_{02} e^{z\beta + x_1\alpha} du\right\} I(u_2 > x_1) \\ &= \lambda^* e^{-\lambda^*(u_2 - x_1)} I(u_2 > x_1), \\ \lambda^* &= \lambda_{02} e^{x_1\alpha + z\beta}. \end{aligned}$$

$U$ 의 밀도함수는 다음과 같다.

$$\begin{aligned} f_U(u_1 | z) &= P(U = u_1, T > u_1) \\ &= \lambda_{02} e^{z\beta} \exp\left\{-\int_0^{u_1} \lambda_{02} e^{z\beta} du\right\} \exp\left\{-\int_0^{u_1} \lambda_{01} e^{z\beta} dx\right\} \\ &= (1 - \pi) \lambda e^{-\lambda u_1} \end{aligned}$$

모의실험에 필요한 표본들은 다음과 같이 발생하여 사용하였다.

- [1]  $Z$ 의 값  $z$ 는  $p=0.5$ 인 베르누이 분포로부터 얻었다.
- [2]  $u$ 의 값은 일양분포  $(0, 1)$ 로 부터 얻었다.
- [3] (a)  $u < \pi$ 이면  $x_1$  값은 모수가  $\lambda$ 인 지수분포로부터 얻고  $u_2$  값은  $f_{U|T}(u_2 | x_1, z)$ 로 부터 얻었다.
- (b)  $u > \pi$ 이면  $u_1$  값은 모수가  $\lambda$ 인 지수분포로부터 얻었다.
- [4]  $C$ 의  $c$  값은 모수가  $\lambda_c$ 인 지수분포로부터 얻었다.

$l(l = 1, \dots, N)$  번째 반복 실험에서  $\theta = (\beta, \alpha)$ 의 최대 편우도 추정량은 뉴턴-랩슨(Newton-Raphson)법을 사용하여 편우도 함수를 최대화하는 것으로 얻어진다.

## 5.2 모의실험의 결과 및 결론

중도절단률이 최대 편우도 추정량에 미치는 영향을 알아보기 위하여 표 3.1에서 제시한 것처럼 발생 가능한 모든 경우를 4개의 범주로 나누었다.

$P_i(z)$ 를  $i$ 번째 범주에 대한  $Z = z$ 가 주어진 조건부확률이라고 하자. 즉,

$$P_i(z) = P(\text{범주 } i \text{에서 나온 결과} | z), \quad i = 1, \dots, 4.$$

$i$ 번째 범주에 대한 비조건부 확률은 다음과 같다.

$$P_i = \sum_z p(z) P_i(z), \quad i = 1, \dots, 4.$$

단,  $p(z)$ 는  $Z$ 에 대한 확률질량함수이다.

범주 1에 대하여만 고려해보면  $P_1$ 은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} P_1 &= \sum_z p(z) P_1(z), \\ P_1(z) &= \Pr(C \leq t, C \leq U) \\ &= \frac{\pi \lambda_c}{\lambda + \lambda_c}. \end{aligned}$$

본 절에서는  $\alpha = \beta = 0.7$ 인 경우 1000회의 모의실험을 통하여 얻은  $\alpha$ 와  $\beta$ 의 최대 편우도 추정량의 히스토그램을 이용하여  $\hat{\alpha}$ 과  $\hat{\beta}$ 의 정규성을 검토해 보았다.

<그림 1>과 <그림 2>에 의하여  $P_3$ 가 가장 큰 경우, 즉 백혈병 환자가 골수이식 수술을 받은 후 임의 중도 절단된 환자들의 비율이 가장 많은 경우와  $P_4$ 가 가장 큰 경우, 즉 골수 이식수술을 받고 난 후 사망하거나 재발한 환자의 비율이 가장 많은 경우  $\hat{\alpha}$ 의 분포가 정규성을 따른다고 볼 수 있으며,  $P_3$ 가 가장 큰 경우보다  $P_4$ 가 가장 큰 경우에  $\hat{\alpha}$ 의 분포가 더 정규성에 가까워짐을 알 수 있다.

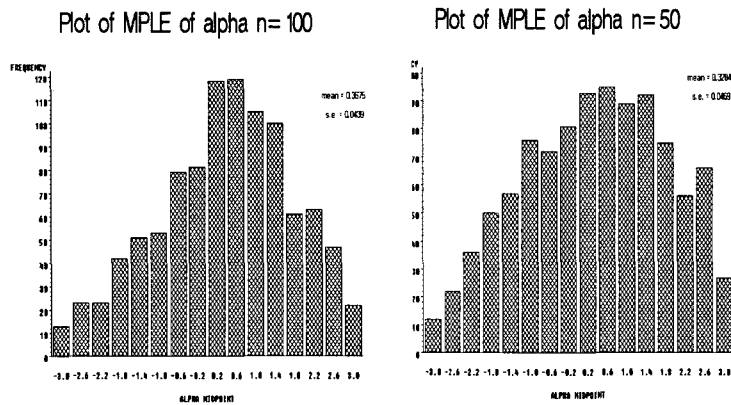
<그림 3>과 <그림 4>에서 알 수 있듯이 백혈병 환자가 골수이식 수술을 받은 후 임의 중도 절단된 환자들의 비율이 가장 많은 경우와 골수 이식수술을 받고 난 후 사망하거나 재발한 환자의 비율이 가장 많은 경우 모두  $\hat{\beta}$ 의 분포도 정규성을 따른다고 볼 수 있으며  $P_3$ 가 가장 큰 경우와  $P_4$ 가 가장 큰 경우 정규성을 따르는 정도에는 큰 차이를 발견할 수 없다.

백혈병 환자가 골수이식 수술을 받거나 사망 전에 임의 중도 절단된 비율이 가장 많은 경우와 골수 이식 수술을 받기 전 재발하거나 사망하여 정보 중도절단 된 비율이 가장 많은 경우에는  $\hat{\beta}$ 의 불편성과 정규성은 보장된다고 볼 수 있으나  $\hat{\alpha}$ 의 정규성은  $n=100$ 인 경우에도 보장된다고 볼 수 없었다.

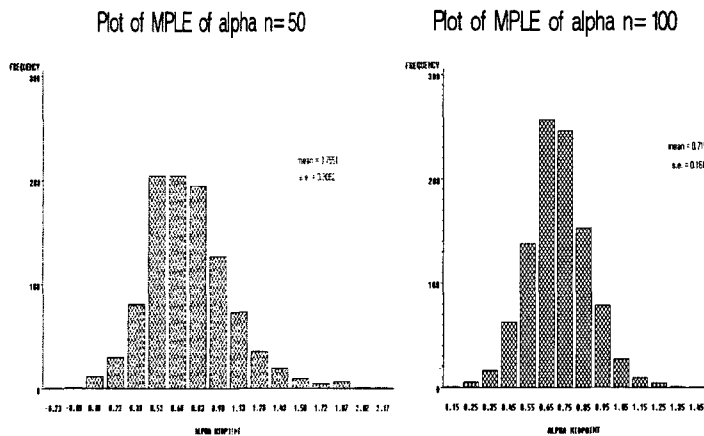
최대 편우도 추정량 방법은 백혈병 환자가 골수 이식 수술을 받고 난 후 재발하거나 사망한 환자의 비율이 다른 세 가지 경우보다 큰 경우에만  $\alpha$ 를 추정하기 위한 효율적인 방법으로 나타났다. 임의적 또는 정보적 중도절단의 비율이 많은 경우와 재발하고 난 후 임의

중도 절단된 환자의 비율이 많은 경우에는 최대 편우도 추정량을 사용하는 것은 바람직하지 않은 것으로 나타났다. 회귀계수  $\beta$ 를 추정하기 위해서는 어떤 경우든지 최대 편우도 추정량 방법을 사용하는 것이 바람직하다는 결론을 얻었다.

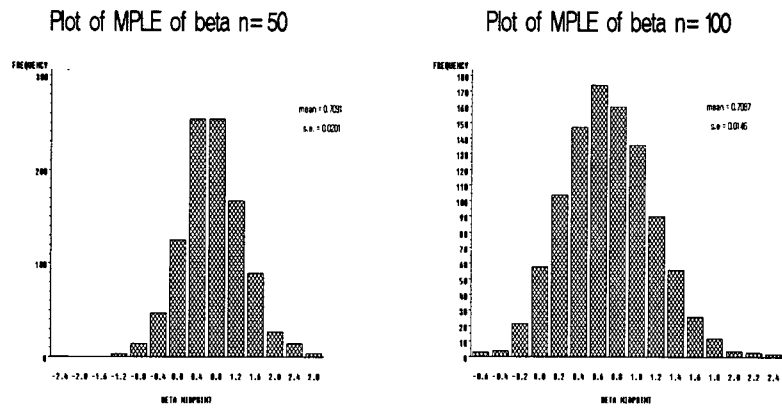
자료 분석시 흔히 범할 수 있는 오류, 즉 정보적 중도 절단 자료를 임의의 중도절단 자료로 잘못 다루었을 경우 최대 편우도 추정 방법의 변화를 고려하여 최대 편우도 추정 방법의 효율성을 고려해보는 것도 추후 중요한 연구 과제라고 볼 수 있다.



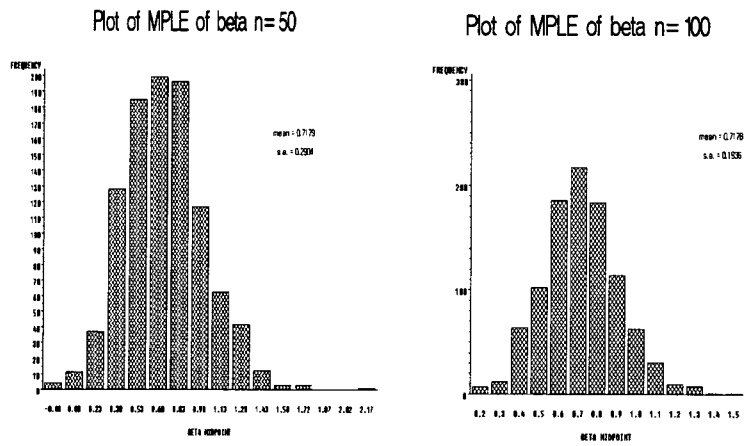
<그림 1>  $P_3$ 가 가장 큰 경우,  $\hat{\alpha}$ 의 그림



<그림2>  $P_4$ 가 가장 큰 경우,  $\hat{\alpha}$ 의 그림



<그림 3>  $P_3$ 가 가장 큰 경우,  $\hat{\beta}$ 의 그림



<그림 4>  $P_4$ 가 가장 큰 경우,  $\hat{\beta}$ 의 그림

## 참고문헌

- [1] Fleming, T. R. & Harrington, D. P.(1991), *Counting Processes and Survival Analysis*, New York, Wiley,.
- [2] Kalbfleisch, J. D. & Prentice, R. L. . (1980), *The Statistical Analysis of Failure Time Data*, New York, Wiley.
- [3] Klein, J. P. ,Keiding, N. & Kamby, C.(1989), Semiparametric Marshall-Olkin Models Applied to the Occurrence of Metastases at Multiple Sites after Breast Cancer, *Biometrics*, 45, 1073-1086.
- [4] Lagakos, S. W(1979), General Right Censoring and Its Impact on the Analysis of Survival Data, *Biometrics*, 35, 139-156.
- [5] Marshall, A. W., & Olkin, I.(1967), A Multivariate Exponential Distribution, *Journal of the American Statistical Association*, 62, 30-44.
- [6] Nádas, A(1971), The Distribution of the Identified Minimum of a Normal Pair Determines the Distribution of the Pair, *Technometrics*, 13, 201-202.
- [7] Smith, F. B. & Helms, R. W(1995), EM Mixed Model Analysis of Data from Informatively Censored Normal Distributions, *Biometrics*, 51, 425-436.