The Korean Communications
in Statistics Vol. 10 No. 2, 2003
pp. 267-275

A Bayesian Approach for Accelerated Failure Time Model with Skewed Normal Error

Chansoo Kim[1)]

Abstract

We consider the Bayesian accelerated failure time model. The error distribution is assigned a skewed normal distribution which is including normal distribution. For noninformative priors of regression coefficients, we show the propriety of posterior distribution. A Markov Chain Monte Carlo algorithm(i.e., Gibbs Sampler) is used to obtain a predictive distribution for a future observation and Bayes estimates of regression coefficients.

Keywords : Accelerated failure time model, Markov Chain Monte Carlo, Skewed normal distribution.

1. 서론

Cox의 비례위험모형은 그 취급성과 유용성으로 인해 생존자료를 모형화 하는데 널리 이용되어진다. 그러나 비례위험모형을 적용하기 어려운 경우 그 대안으로 가속화 고장시간 모형을 고려할 수 있다.

생존시간(고장시간)을 T라고 하고, 이 생존시간에 영향을 주는 공변량을 X라 할 때, 가속화 고장시간 모형(accelerated failure time model)은

$$T = \exp(-\beta' X) V. \tag{1.1}$$

여기서 V는 공변량 X의 값이 0일때의 생존시간이고, 식 (1.1)은 생존시간 T가 X값에 따라 지수적으로 변화하는 것을 의미한다.

생존시간 $T_1, \cdots, T_n$이 가속화 고장시간 모형으로부터 나왔다고 할 때, 모형 (1.1)식 양변에 자연로그를 취하면 다음과 같은 로그-선형모형을 얻을 수 있다

$$Y_i = \log T_i = -\beta' X_i + \Theta_i, \quad i = 1, \cdots, n. \tag{1.2}$$

여기서 $X_i = (x_{i1}, \cdots, x_{ip})$는 i번째 개체에 대한 설명변수들의 벡터이고 β는 p개의 미지의 회귀계수 벡터이고 $\Theta_i = \log V_i$는 오차항이다. (1.2)식과 같은 선형회귀모형에서 오차항은 정규분포로 가정하나 많은 실제적인 문제들에서 비 정규분포를 갖는 오차항이 더 유용하다.따라서, 본 논문에서

1) Full-time Lecturer, Department of Applied Mathematics, Kongju National University, Kongju, 314-702
chanskim@kongju.ac.kr

는 오차항이 비 정규분포중의 하나인 왜도 정규분포를 갖는 경우 베이지안 접근 방법을 제시하고자 한다.

가속화 고장시간 모형에 대한 고전적 해석은 Kalbfleisch와 Prentice(1980)에 의해 논의되어졌다. 베이지안 접근방법으로 크게 두 가지의 측면에서 고려할 수 있는데, 먼저 이 모형에 대한 베이지안 준모수적(Bayesian semi-parametric) 접근 방법은 Christensen과 Johson(1988)과 Johnson과 Christensen(1989)에 의해 전개되어졌다. 그들은 오차항 V_i가 Ferguson(1973)이 제시한 디리클릿 과정(Dirichlet process)을 가정하고 회귀계수와 생존율을 추정하였다. 그러나 디릿클릿 과정의 이산적 성질에 기인되는 문제와 비중도 절단된 경우에 베이지안 해석에 어려움을 가지고 있다. 따라서, 디릿클릿 과정을 이용하는 대신 최근에 Walker와 Mallick(1999)이 디릿클릿 과정의 확장된 개념으로 오차항에 Polya tree 사전분포를 가정하고 생존함수를 추정하였다.

또, 다른 접근 방법으로 비 정규 오차항을 갖는 모형을 고려할 수 있다. 최초로Zellner(1976)는 다변량 스튜던트 t 분포를 이용한 선형회귀모형을 베이즈 방법과 고전적 방법으로 연구, 비교하였고 Azzalini와 Dalla-Valle(1996)은 정규분포에 모양 모수(shape parameter)를 추가한 족을 확장시킨 왜도 정규분포의 다변량 형태에 대한 일반적인 이론을 제시했다. 최근에 Branco와 Dey(2001)는 다변량 정규분포, 스튜던트 t분포, 멱 지수분포와 피어슨 타입 II분포를 포함하는 다변량 왜도 타원형 분포의 일반적인 족을 제안하면서 왜도 모수를 정의하였으며, Sahu, Dey와 Branco(2001)는 왜도 타원형 오차를 갖는 회귀모형을 연구하여 회귀계수에 대한 베이지안 방법을 제안하였다.

본 논문에서는 Chen, Dey와 Shao(1999)가 제시한 비 대칭 확률분포를 이용하여, 오차항이 왜도 정규분포를 갖은 가속화 고장시간 모형에 대한 베이지안 접근 방법을 제시하고자 한다. 먼저 사전분포가 비 적절한(improper) 경우, 사후분포의 적절성을 보이고자 한다. 또한 계산상의 어려움을 해결하기 위해 MCMC 기법을 사용하여, 회귀계수와 생존함수에 대한 추정 문제를 다루고자 한다. 마지막으로 오차항이 정규분포와 왜도 정규분포 중 어느 모형이 타당한지를 베이지안 슈와르쯔 정보 기준을 적용하여 모형선택을 하고 모형의 타당성을 살펴보기 위해, Feigel와 Zelen(1965)에의해 분석된 백혈병환자 자료에 적용하고자 한다.

2. 왜도 정규분포

이장에서는 왜도 정규분포에 대해 소개하고, 가속화 고장시간 모형에 대한 베이지안 방법을 적용하고자 한다. 먼저, 오차항이 서로 독립이고 왜도 정규분포(skewed normal distribution)를 갖는다고 가정하자. 그러면, Chen, Dey와 Shao(1999)는 다음과 같은 비대칭 확률변수를 제시하였다.

$$\Theta = \delta Z + U. \tag{2.1}$$

여기서 U은 대칭이고 단봉(unimodal)을 갖는 분포이고 Z는 양수이고 비대칭 분포를 가진다. 이때 $\delta = 0$이면 Θ 은 일반적인 대칭인 형태를 갖는 분포를 가진다. 만약 $\delta > 0 (\delta < 0)$이면 오른쪽(왼쪽)으로 왜도를 갖는 모형을 갖는다. 따라서 δ 는 왜도 모수(skewness parameter)로서 해석할 수 있다. 이러한 성질을 살펴보기 위해, Z와 ε이 3차 적률을 갖는다고 하자. σ_Z^2와 σ_ε^2을 각각 Z와 ε의 분산이고 μ_Z^3 을 Z의 표준화된 3차 적률(standardized third moment)이라고 하자. 즉 $\mu_Z^3 = E\left(\frac{Z - E(Z)}{\sigma_Z}\right)^3$. 그러면, W 의 표준화된 3차 적률은 다음과 같다.

$$\mu_W^3 = E\left(\frac{W - E(W)}{\sigma_W}\right)^3 = \frac{\delta^3\sigma_Z^3\mu_Z^3}{\sigma_W^3}.$$

여기서, $\sigma_W^2 = Var(W) = \delta^2\sigma_Z^2 + \sigma_\varepsilon^2$이다. 만약 Z 오른쪽으로 치우쳐 있다는 것은 W의 주변분포는 $\delta > 0 (\delta < 0)$일 때 오른쪽(왼쪽)으로 왜도를 갖는다는 것을 볼 수 있다.

(2.1)식에서 확률변수 Z와 U가 각각 절단된 표준정규분포와 정규분포를 갖는다면, Θ는 왜도 정규분포를 따른다. 본 논문에서는 Chen, Dey와 Shao(1999)가 제시한 비대칭 분포의 특별한 형태인 왜도 정규분포를 가속화 고장시간 모형에 적용하고자 한다. 따라서 모형 (1.2)의 오차항에 (2.1)식을 결합하면 다음과 같은 모형이 된다.

$$\begin{aligned} Y_i &= -X_i\beta + \sigma\delta(Z_i - c) + \sigma U_i, \quad i = 1, \cdots, n, \\ Z_i &\sim N^+(0,1), \quad U_i \sim N(0,1). \end{aligned} \tag{2.2}$$

여기서, N^+은 절단된 표준정규분포를 나타내고 $c = E(Z_i) = \sqrt{(2/\pi)}$. 따라서 σ, δ, β가 주어질 때, $(y_i\ Z_i)$의 결합확률분포는 다음과 같이 표현된다.

$$[y_i, z_i | \sigma, \delta, \beta\ x] = N(y_i - x_i'\beta + \sigma\delta(z_i - c), \sigma^2)N^+(z_i : 0, 1), \quad i = 1, \cdots, n.$$

2.1. 비정보적 사전분포

σ, δ와 β에 대해 비정보적 사전 분포(noniformative prior), 즉 $\pi(\beta, \sigma^2, \delta) \propto \pi(\sigma^2)\pi(\delta)$ 를 갖는다고 하자. 여기서 $\beta = (\beta_1, \cdots, \beta_p)$. 비정보적 사전 분포, σ, δ와 β들이 서로 독립임을 가정하고 각각의 사전분포들은 다음과 같은 분포를 갖는다. $\pi(\beta) \propto 1$이고 $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$ 이고 $\pi(\delta) \propto \exp\left(-\frac{(\delta - \mu_\delta)^2}{2\sigma_\delta^2}\right)$이다.

이와같은 비모수적 사전분포로부터, 관측된 자료 Y를 근거로 한 사후분포는 다음과 같이 주어진다.

$$\begin{aligned} [\beta, \delta, \sigma^2, Z | y] \propto &\left(\frac{1}{2\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (y_i + x_i'\beta - \sigma\delta(z_i - c))^2}{2\sigma^2}\right) \\ &\times \prod_{i=1}^n \exp\left(-\frac{z_i^2}{2}\right) 1(z_i > 0)\, \pi(\sigma^2)\pi(\beta)\pi(\delta). \end{aligned} \tag{2.3}$$

사전분포가 비적절(improper)이므로 사후분포에 대한 적절성(propriety)에 대한 논의가 필요하다. 따라서 아래의 정리는 사후분포가 적절하기 위한 조건을 제시하고 있다.

Theorem 2.1. X^*가 행 x_i'을 갖는 $n \times p$ 행렬이라 하자. 만약 (a) X^*는 풀 랭크를 갖고 (b) $\pi(\delta)$는 적절하고 (c) $n > p$이면, (2.3)에 주어진 사후분포는 적절하다.

Proof. (2.3)이 적절하다는 것을 보이기 위해, 정규화 상수가 유한하다는 것을 보이는 것을 보이면 된다. 즉,

$$Q = \int_\delta \int_Z \int_{\sigma^2} \int_\beta \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(y_i + x_i'\beta - \sigma\delta(z_i - c))^2}{2\sigma^2}\right)$$
$$\times \prod_{i=1}^{n} \exp\left(-\frac{z_i^2}{2}\right) 1(z_i > 0) \frac{1}{\sigma^2}\, \pi(\delta)\, d\beta\, d\sigma\, dZ\, d\delta,$$

$$= \int_\delta \int_Z \int_{\sigma^2} \int_\beta \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}[\beta + (X^{*\prime}X^*)^{-1}X^{*\prime}(\boldsymbol{y} - \sigma\delta Z^*)]'\right.$$
$$\left.(W'W)[\beta + (X^{*\prime}X^*)^{-1}X^{*\prime}(\boldsymbol{y} - \sigma\delta Z^*)]\right)$$
$$\times \exp\left(-\frac{1}{2} Z'Z\right) \frac{1}{\sigma^2}\, \pi(\delta)\, d\beta\, d\sigma\, dZ\, d\delta.$$

여기서, $X^* = [x_{ij}]_{i=1,\cdots,n}^{j=1,\cdots,p}$ 는 $n \times p$ 행렬이고 Z^*는 $z_i - c$를 원소로 갖는 $n \times 1$ 벡터이고 $i = 1, \cdots, n$. 따라서, Q가 유한하다는 것을 보이면 되므로 Q 는 다음과 같이 표현될 수 있다.

$$Q = \int_\delta \int_Z \int_{\sigma^2} \left(\int_\beta \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{p}{2}} |X^{*\prime}X^*|^{-1/2} \right.$$
$$\times \exp\left(-\frac{1}{2\sigma^2}[\beta + (X^{*\prime}X^*)^{-1}X^{*\prime}(y - \sigma\delta Z^*)]'(X^{*\prime}X^*)\right.$$
$$\left.\left.[\beta + (X^{*\prime}X^*)^{-1}X^{*\prime}(y - \sigma\delta Z^*)]\right) d\beta\right)$$
$$\times \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n-p}{2}} |X^{*\prime}X^*|^{1/2} \exp\left(-\frac{1}{2} Z'Z\right) \frac{1}{\sigma^2}\, \pi(\delta)\, d\sigma\, dZ\, d\delta,$$

$$= \int_\delta \int_Z \int_{\sigma^2} \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n-p}{2}} |X^{*\prime}X^*|^{1/2} \exp\left(-\frac{1}{2} Z'Z\right) \frac{1}{\sigma^2}\, \pi(\delta)\, d\sigma^2\, dZ\, d\delta,$$

$$= |X^{*\prime}X^*|^{1/2} \int_{\sigma^2} \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n-p}{2}} \frac{1}{\sigma^2} \left(\int_Z \exp\left(-\frac{1}{2} Z'Z\right) dZ \right) \left(\int_\delta \pi(\delta)\, d\delta \right) d\sigma,$$

$$= |X^{*\prime}X^*|^{1/2} 2^{-n} (2\pi)^{\frac{p}{n}} \int_{\sigma^2} \left(\frac{1}{\sigma^2}\right)^{\frac{n-p}{2}} \frac{1}{\sigma^2}\, d\sigma.$$

만약에 $n > p$이면 마지막 적분값은 유한하다. 따라서 조건 (a), (b)와 (c)하에서, 사후분포는 적절하다.

2.2. 정보적 사전분포

모수 $(\beta, \sigma^2, \delta)$에 대해, 다음과 같은 형태를 갖는 정보적 사전분포(informative prior)를 가정하자. 즉, $[\beta_i] = N(\mu_\beta, \sigma_\beta)$, $i=1,\cdots,p$, $[\delta] = N(\mu_\delta, \sigma_\delta^2)$이고 $[\sigma^2] = IG(a, b)$이다. 결합사후분포는 우도함수와 사전분포들의 곱에 비례하므로 결합사후분포는 (2.3)식과 같다.

(2.3)식으로부터 $(\beta, \sigma^2, \delta)$ 각각에 대한 주변사후분포(marginal posterior distribution)을 직접적으로 얻기가 어렵기 때문에, Gibbs sampler를 적용하여 모수들에 대한 추정을 하고자한다. 그러면, 결합사후분포 (2.3)로부터 완전조건부 분포(full conditional distributions, FCD)들은 다음과 같이 쉽게 얻어진다.

(1) $[\beta_j|\beta_{(j)}, \delta, \sigma^2, Z, Y] = N\left(\frac{A}{B}, \frac{1}{B}\right)$, $j=1,\cdots,p$,

(2) $[\delta|\beta_1,\cdots,\beta_p, \sigma^2, Z, Y] = N\left(\frac{C}{D}, \frac{1}{D}\right)$,

(3) $[Z_i|\beta_1,\cdots,\beta_p, \delta, \sigma^2, Z_{(i)}, Y] = N^+\left(\frac{\delta(y_i + x_i'\beta + c\delta\sigma)}{\sigma(1+\delta^2)}, \frac{1}{1+\delta^2}\right)$,

(4) $[\sigma^2|\beta_1,\cdots,\beta_p, \delta, Z, Y] \propto IG\left(\frac{n}{2}+a, \frac{\sum_{i=1}^n (y_i + x_i'\beta)^2}{2} + b\right)$

$$\times \exp\left(\frac{\delta\sum_{i=1}^n (z_i - c)(y_i + x_i'\beta)}{\sigma}\right).$$

여기서

$$A = -\frac{\sum_{i=1}^n x_{ij}(y_i + x_{i(j)}'\beta_{i(j)} - \sigma\delta(z_i - c))}{\sigma^2} + \frac{\mu_\beta}{\sigma_\beta^2}, \quad B = \frac{\sum_{i=1}^n x_{ij}^2}{\sigma^2} + \frac{1}{\sigma_\beta^2},$$

$$C = \frac{\sum_{i=1}^n (z_i - c)(y_i + x_i'\beta)}{\sigma} + \frac{\mu_\delta}{\sigma_\delta^2}, \quad D = \sum_{i=1}^n (z_i - c)^2 + \frac{1}{\sigma_\delta^2}.$$

여기서 완전조건분포 (1), (2)와 (3)은 쉽게 샘플링 할 수 있다. 그러나, σ^2의 FCD는 쉽고 명확한 형태를 갖고 있지 않기 때문에 직접적으로 난수를 생성하기 어렵다. 따라서 Metropolis 알고리즘을 이용하여 계산상의 어려움을 해결하고자한다. 먼저 $\tau^2 = \log\sigma^2$라 하자. 그러면 $\beta_1,\cdots,\beta_p, \delta, Z, Y$ 가 주어질 때, τ^2 의 조건부 분포는 다음과 같이 변환되어진다.

$$[\tau^2|\beta_1,\cdots,\beta_p, \delta, Z, Y] \propto \exp\left(-\frac{n+2a}{2}\tau^2 - e^{-\tau^2}\left(\frac{\sum_{i=1}^n (y_i + x_i'\beta)^2}{2} + b\right)\right.$$

$$\left. + e^{-\frac{\tau^2}{2}}\delta\sum_{i=1}^n (z_i - c)(y_i + x_i'\beta)\right). \qquad (2.4)$$

τ^2를 생성하기 위해, 정규커널 $N(\widehat{\tau^2}, \widehat{\sigma^2_{\tau^2}})$을 사용한다. 여기서 $\widehat{\tau^2}$은 (2.4)식의 오른쪽에 자연로그를 취한 후 최대로 되는 값이고 이 값은 O'Neill(1971)에 의해 활용되어진 Nelder-Mead 알고리즘을 사용하여 얻어진다. 또한, $\widehat{\sigma^2_{\tau^2}} = -\frac{d^2 \log[\tau^2|\beta, \delta, Z, Y]}{d(\tau^2)^2}\Big|_{\tau^2=\widehat{\tau^2}}$. 따라서 τ^2을 생성하기 위한 알고리즘은 다음과 같다.

단계 1. τ_k^2 을 현재 값이라고 하자.

단계 2. τ_k^{*2}을 $N(\widehat{\tau^2}, \widehat{\sigma^2_{\tau^2}})$부터 생성한다.

단계 3. τ_k^2 에서 τ_k^{*2}로 채택될 확률 α은 다음과 같다.

$$\alpha = \min\left\{1, \frac{P(\tau_k^{*2}|\beta,\delta,z,y)\phi(\frac{\tau_k^2 - \widehat{\tau_k^2}}{\widehat{\sigma^2_{\tau^2}}})}{P(\tau_k^2|\beta,\delta,z,y)\phi(\frac{\tau_k^{*2} - \widehat{\tau_k^2}}{\widehat{\sigma^2_{\tau^2}}})}\right\}.$$

3. 모형선택과 생존함수 추정

여기서 공변량 $\boldsymbol{X}$가 주어질 때 새로운 환자에 대한 생존함수에 대해 예측하고, 오차항이 정규분포와 왜도 정규분포 중에서 어느 모형을 갖는 것이 타당한지에 대한 모형선택을 하고자한다. 먼저 왜도에 대한 효과가 있는지를 알아보기 위해 다음과 같은 두 모형을 고려하고자 한다.

$$M_1: Y_i = -X_i\beta + U_i, U_i \sim N(0, \sigma^2) \quad vs \quad M_2: \text{모형 (2.2) 을 갖는다} \qquad (3.1)$$

여기서 (2.2)식에서 δ=0이면 모형 M_1을 갖고 그렇지 않으면 M_2 를 갖는다는 것을 알수 있다. 따라서 모형 M_1은 오차항이 정규분포를 따르고, 모형 M_2은 오차항이 왜도 정규분포를 따른다. 모형의 선택 기준으로 베이지안 슈와르쯔 정보 기준(Bayesian Schwartz information criterion, BSIC)을 사용한다. 원래 슈와르쯔 정보기준(Schwartz information criterion, SIC)은 $-2\ \log \mathrm{L}(\hat{\theta}|\mathrm{D}) + \mathrm{p}\ \log \mathrm{n}$와 같이 정의 된다. 여기서 $L(\hat{\theta})$은 모형에 대한 우도함수이고, $\hat{\theta}$은 미지의 모수들에 대한 최대우도추정량이고, p는 추정되어질 미지의 모수들의 개수이고 n은 표본의 크기이다. BSIC은 Yang과 Kuo(2000)에 제시한 다음과 같은 형태이다.

$$\mathrm{BSIC} = \int(-2\ \log \mathrm{L}(\theta|\mathrm{y}_1, \cdots, \mathrm{y}_n) + \mathrm{p}\ \log \mathrm{n})[\theta|\mathrm{y}_1, \cdots, \mathrm{y}_n]\mathrm{d}\theta.$$

여기서 $L(\theta|D)$ 는 우도함수이고, p 추정되어질 미지의 모수의 개수이고 $[\theta|y_1, \cdots, y_n]$은 사후분포를 나타낸다. BSIC의 계산은 사후분포에 대한 적분값이므로 계산이 용이하지 않다. 따라서 BSIC의 추정치로서 다음과 같은 몬테 카를로 추정치를 계산할 수 있다.

$$\widehat{\mathrm{BSIC}} = \frac{1}{T}\sum_{t=1}^{T} -2\ \log L(\theta^{(t)}|y_1, \cdots, y_n) + p_1\ \log n.$$

여기서 T 총 반복횟수이고 $\theta^{(t)}$은 깁스 샘플러(Gibbs sampler)의 결과 값이다. 따라서, 모형선택

에 대한 판정은 만약에 $\widehat{BSIC}_1 \leq \widehat{BSIC}_2$ 이면, 모형 M_1을 선택하고 그렇지 않은 경우에는 모형 M_2을 선택한다.

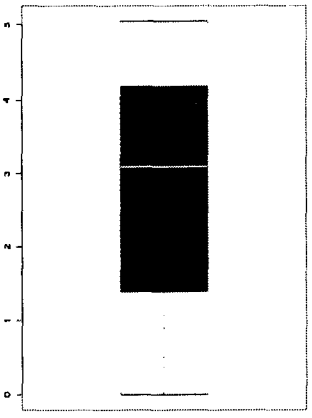
또한, 새로운 환자에 대해서 생존함수를 예측하는 문제를 다루고자한다. 추정하고자 하는 관심은 임의의 $t>0$ 대해 $P(T>t|data)=1-P(T\leq t|data)$이다. $\Psi=(\beta,\delta,\sigma^2,Z)$라 하면, $P(T\leq t|data)=\int P(T\leq t|data,\Psi)P(\Psi|data)d\Psi$ 이고 깁스 샘플러로부터 얻어진 표본 $\Psi^{(l)}=(\beta^{(l)},\delta^{(l)},\sigma^{2(l)},Z^{(l)})$을 이용하여 $\frac{1}{L}\sum_{l=1}^{L}P(T\leq t|data,\Psi^{(l)})$와 같이 계산할 수 있다. 여기서 L 은 생성된 샘플들의 총 반복 횟수이다. 따라서 생존함수에 대한 추정은 다음과 같이 구해진다.

$$\hat{S}(t)=1-\frac{1}{L}\sum_{l=1}^{L}P\left(U\leq\frac{\log t+X\beta^{(l)}-\sigma^{(l)}\delta^{(l)}(Z^{(l)}-c)}{\sigma^{(l)}}\,|\,data,\Psi^{(l)}\right).$$

4. 예제

이장에서는 앞에서 제시한 모형의 타당성을 알아보기 위해 Feigl 과 Zelen(1965)이 분석한 다음과 같은 백혈병 환자에 대한 자료를 적용하고자 한다. AG(Auer rods and/or graulature) 양성군과 AG 음성군으로 구성된 33명의 백혈병 환자의 생존시간(단위 : 주(week))을 비교하고자 한다. 여기서 공변량 x_{i1}=1, x_{i2}는 AG 양성군이면 0, AG음성군이면 1 이고 x_{i3}는 백혈구수치(white blood cell count)이다. 관측된 환자의 생존시간에 자연대수를 취한후, 상자그림을 살펴보면 다음과 같다.

그림 4.1. 환자의 생존시간에 대한 상자그림



생존시간은 비대칭이고 왼쪽으로 왜도를 갖는 형태임을 알 수 있다. 따라서, 왜도 정규분포 모형을 가정하는 것이 바람직하게 보인다. 표 4.1에는 모수들에 대한 베이즈 추정치들이 주어져 있다.

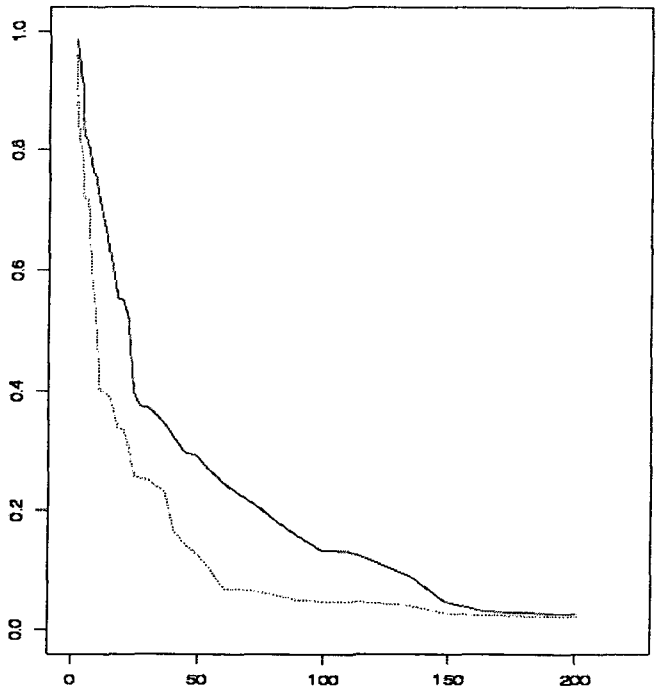
표 4.1. β 와 δ들에 대한 베이즈 추정치

모수	사후평균 (posterior mean)	사후중앙 (posterior median)	표준편차 (standard deviation)
β_1	-1.859	-1.916	1.049
β_2	0.689	0.675	0.473
β_3	-0.123	-0.110	0.116
δ	-0.559	-0.515	0.872

표 4.1에서 볼 수 있듯이 δ의 베이즈 추정치가 -0.559이고, 음수값이므로 왼쪽으로 왜도(left skewness)를 갖는다는 것을 보여주고 있다. 또한 3장에서 주어진 모형선택에서 모형 M_1에서 BSIC의 추정값 $\widehat{\text{BSIC}}_1$ =138.78이고 모형 M_2에서 $\widehat{\text{BSIC}}_2$=125.51이므로 판정기준에 따라, 모형 M_2가 선택되어진다. 즉, 오차항이 정규분포를 갖는 모형보다는 왜도 정규분포를 갖는 모형이 타당하다는 것을 볼 수 있다.

그림 4.3 에서는 예측된 생존함수를 나타내고 있다. 직선은 AG 양성군에 대한 생존함수이고 점선은 AG 음성군에 대한 생존함수 이다. 기존의 연구 결과와 같이 두 집단간에 생존율에 차이가 있음을 알 수 있다.

그림 4.2. AG 양성군과 음성군에 대한 생존함수



References

[1] Azzalini, A. and Dalla Vale, A. (1996) The multivariate skew-normal distribution, *Biometrika*, 83, 715-726

[2] Branco D. and Dey, D. K. (2000) Bayesian regression model under skewed heavy tailed error distribution, Technical Report, Department of Statistics, University of Connecticut.

[3] Chen, M.-H., Dey, D. K. and Shao, Q.-M. (1999) A new skewed link model for dichotomous quantal response data, *Journal of the American Statistical Association*, 94, 1172-1186.

[4] Christensen, R. and Johnson, W. (1988) Modelling accelerated failure time with a Dirichlet process, *Biometrika*, 75, 693-704.

[5] Feigl, P. and Zelen, M. (1965) Estimation of exponential probabilities with concomitant information, *Biometrics*, 21, 826-838.

[6] Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems, *The Annals of Statistics*, 1, 209-230.

[7] Johnson W. and Christensen, R. (1989) Nonparametric Bayesian analysis of the accelerated failure time model, *Statistics and Probability Letters*, 8, 179-184.

[8] Kalbfleisch, J. D. and Prentice, R. L. (1980) *The Statistical Analysis of Failure Time Data*, New York: Wiley.

[9] Nelder, J. A. and Mead, R. (1965) A simplex method for function minimization, *Computer J.*, 7, 308-313.

[10] O'Neill, R. (1971) Algorithm AS47-function minimization using a simplex procedure, *Applied Statistics*, 20, 338-345.

[11] Sahu, S. K., Dey, D. K. and Branco, M. D.(2001) A new class of multivariate skew distributions with applications to Bayesian regression models, Technical Report , University of Southampton.

[12] Walker, S. and Mallick, B. K. (1999) A Bayesian semiparametric accelerated failure time model, *Biometrics*, 55, 477-483.

[13] Yang, T. and Kuo, L.(2000) Bayesian binary segmentation procedure for homogeneous Poisson process with multiple change positions, Technical Report, Department of Statistics, University of Connecticut.

[14] Zellner, A. (1976) Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms, *Journal of the American Statistical Association*, 71, 400-405.

[2003년 1월 접수, 2003년 5월 채택]