비동질적 포아송과정을 사용한 소프트웨어 신뢰 성장모형에 대한 베이지안 신뢰성 분석에 관한 연구

이 상 식† · 김 희 철† · 송 영 재††

요 약

본 논문에서는 비동질 포아송 과정(NHPP)에 기초한 소프트웨어 에러 현상에 대한 신뢰도 모형을 고려하고 사전정보(Prior information)를 이용한 베이지안 추론을 시행하였다. 고장 패턴은 NHPP에 대한 강도함수와 평균값 함수로서 나타낼 수 있다. 따라서 본 논문에서는 대수형 포아송 실행시간 모형(Logarithmic Poisson model), Crow 모형 그리고 Rayleigh 모형에 대하여 베이지안 모수 추정방법을 적용하였다. 효율적 모형을 위하여 이들 모형에 관한 모형선택을 편차자승합(SSE)의 합을 이용하여 시행하였고 모수의 추정을 위해서 마코브체인 몬테카를로(MCMC) 기법중에 하나인 깁스샘플링(Gibbs sampling)과 메트로폴리스 알고리즘을 이용한 근사추정 기법이 사용되었다. 수치적인 예에서는 Musa의 T1 자료를 이용하여 모수 및 신뢰도를 추정한 수치 결과를 나열하였다.

The Bayesian Analysis for Software Reliability Models Based on NHPP

Sangsik Lee† · Heecheul Kim† · Yongjae Kim††

ABSTRACT

This paper presents a stochastic model for the software failure phenomenon based on a nonhomogeneous Poisson process (NHPP) and performs Bayesian inference using prior imformation. The failure process is analyzed to develop a suitable mean value function for the NHPP ; expressions are given for several performance measure. The parametric inferences of the model using Logarithmic Poisson model, Crow model and Rayleigh model is discussed. Bayesian computation and model selection using the sum of squared errors. The numerical results of this models are applied to real software failure data. Tools of parameter inference was used method of Gibbs sampling and Metropolis algorithm. The numerical example by T1 data (Musa) was illustrated.

키워드 : 소프트웨어 신뢰도 모형(Software Reliability Model), 깁스 샘플링(Gibbs Sampling), 비동질적인 포아송 과정(Nonhomogeneous Poisson Process), 대수형 포아송 모형(Logarithmic Poisson Model), 메트로폴리스 알고리즘, 소프트웨어 신뢰도(Software Reliability), 편차자승합(Sum of the Squared Errors).

1. 서 론

신뢰도 이론은 소프트웨어 시스템 전체가 규정된 환경조건하에서 의도하는 기간동안에 요구된 기능을 만족스럽게 수행할 수 있는 확률을 예측하고 증대시키기 위한 실제적인 도구가 된다. 제품이 의도된 기간동안 고장없이 제 기능을 발휘할 수 있는 확률을 보다 정확한 방법으로 추정하는 연구는 이 분야의 관심사항이 된다.

신뢰도의 정량적인 값을 얻는데 있어 기존의 방법은 최우추정법(MLE)을 많이 사용하였다. 그러나 새로운 자료가 얻어지면 그 자료를 이제까지 얻었던 자료와 결합시켜 새로운 결론에 도달하려는 이론이 베이즈 추정법이다. 즉, 베이즈 이론은 알려져 있는 사실에 대한 주관적 의견을 경험이나 지식을 바탕으로 하여 사전정보를 만든 다음 실험을 통하여 얻어진 자료와 결합시켜 사후정보를 추출하는 과정이다. 그러나 베이즈 추정법에서 사전확률 분포인 수명분포가 복잡하면 적분이 불가능해 지기 때문에 사후정보의 추출이 불가능해진다.

따라서 본 연구에서는 최우추정법과 적분이 난해한 경우에 깁스 샘플링(Gibbs sampling)을 이용하여 근사적 깁스 추정량을 유도하여 비교하고 그 특징을 분석하고자 한다

본 논문에서는 참고문헌[2]에서 제시하지 않았던 대수형 포아송 모형(Logarithmic Poisson)과 Crow 모형을 적용하여 모수을 추정하고 베이즈 추정치들을 계산하기 위한 사후분포의 형태를 구하는데 메트로폴리스(Metropolis) 알고리즘을 제안하였다. 사후 분포가 임의의 분포를 따르지 않을 경우는 메트로폴리스 알고리즘을 사용할 수 있는데 이 기법은 깁스 알고리즘의 대안으로 이용된다[4, 8].

깁스 알고리즘은 MCMC(Markov Chain Monte Carlo) 알

† 정 회 원 : 송호대학 정보산업계열 교수
†† 종신회원 : 경희대학교 컴퓨터공학과 교수
논문접수 : 2003년 5월 20일, 심사완료 : 2003년 7월 21일

고리즘이다. 이는 바람직한 사후 분포로서 정상성 분포를 가지는 마코프 체인에 따라 표본들이 변화하는 알고리즘이다[4,7,8]. 이 마코프 체인의 추이 측정은 주로 보통 조건부 밀도함수의 곱이다. 즉, 한 변수는 소프트웨어에 남아있는 수많은 오류들이며, 다른 변수는 포아송 확률들 중에서 척도모수(Scale parameter)의 변화에 따라서 어떤 요소가 도움을 주는지를 나타내는 것이다.

본 논문의 구성의 2장에서는 관련 연구로서 소프트웨어 신뢰도에 대한 일반적 내용과 대수형 포아송 실행 시간 모형, Crew 모형 그리고 Rayleigh 모형[13]에 대하여 요약하였고, 3장에서는 깁스 알고리즘을 이용한 모수 추정에 대하여 서술 하였고, 4장에서는 모형선택에 관한 문제를 요약하였고 5장에서는 수치 자료인 T1 자료[13]을 이용하여 각 모형에 대한 모수 추정 및 모형 비교에 대한 수치적인 결과를 마지막으로 6장에서는 결론 및 향후 연구방향을 제시하였다.

2. NHPP 모형과 소프트웨어 신뢰도

어떤 시간 t까지에 발견된 총 에러수를 나타내는 계수과정 $\{N(t), t \geq 0\}$을 고려하자. 이 확률 과정에 대하여 단위 시간당 발견되는 에러수를 나타내는 강도함수 $\lambda(t)$를 가지는 비동질 포아송 과정(NHPP)을 따른다면 테스트 공정 혹은 운용단계에 있어서 에러 발견사상에 대하여 다음의 가정을 만족함이 알려져 있다.

가정 1 : $N(0) = 0$. 즉, 테스트 시각 $t = 0$에서는 에러가 발견되지 않는다.

가정 2 : $\{N(t), t \geq 0\}$은 독립증분을 갖는다. 즉, 서로 다른 테스트 구간(시간)에서 발견 되는 총 에러수는 통계적으로 독립이다.

가정 3 : $\Pr\{N(t+\triangle t) - N(t) = 1\} = \lambda(t)\triangle t + o(\triangle t)$. 즉, 임의의 극소 구간(시간) $\triangle t$에서 한 개의 에러가 발견될 확률은 강도함수 $\lambda(t)$에 비례 한다.

가정 4 : $\Pr\{N(t+\triangle t) - N(t) \geq 2\} = o(\triangle t)$. 즉, 임의의 극소 테스트 구간(시간) $\triangle t$ 에서 두개 이상의 에러가 발견될 확률은 무시해도 좋을 만큼 적다.

위의 가정하에서 $N(t)$를 시간 $(0, t]$사이에 발견된 에러수라 정의하면 $N(t)$는 평균 값 함수(Mean value function) $m(t) = E[N(t)]$인 NHPP에 의해 다음과 같이 모형화 될 수 있다[13, 14].

$$P(N(t) = n) = \frac{m(t)^n}{n!} e^{-m(t)} \quad n = 0, 1, 2, \cdots \tag{2.1}$$

단,

$$m(t) = \int_0^t \lambda(x)dx \tag{2.2}$$

$m(t)$가 t에 대한 비감소 함수(Nondecreasing function) 추세를 가진 미분가능 함수이면 강도 함수(Intensity function)는 $\lambda(t) = m'(t)$가 됨이 알려져 있다. $\lambda(t)$가 상수($m(t)$가 선형(Linear) 추세)이면 동질적 포아송 과정이고, t에 대한 함수형태이면 NHPP가 된다.

2.1 대수형 포아송 모형

본 논문에서는 Okumoto[15]가 정의한 대수형 모형에 기초한 NHPP를 이용하고자 한다. 이 모형에 대한 발생확률인 강도함수는 다음과 같이 정의하였다.

$$\lambda_1(t) = \lambda_0 e^{-\theta m_1(t)} \quad (\lambda_0 > 0,\ \theta > 0) \tag{2.3}$$

단, λ_0는 초기 고장강도를 의미하고 또, θ는 소프트웨어 고장 1개당 고장강도 $\lambda_1(t)$의 감소율을 의미한다. 따라서 식 (2.3)은

$$\frac{dm_1(t)}{dt} = \lambda_0 \exp\{-\theta m_1(t)\} \tag{2.4}$$

인 것으로부터

$$\frac{dm_1(t)}{dt} \exp[\theta m_1(t)] = \lambda_0 \tag{2.5}$$

로 된다. 그리고 $\exp[\theta m_1(t)]$를 t에 관하여 적분하면 다음과 같은 관계식을 얻는다.

$$\frac{d\exp[\theta m_1(t)]}{dt} = \theta \frac{dm_1(t)}{dt} \exp[\theta m_1(t)] \tag{2.6}$$

식 (2.5)과 식 (2.6)의 관계를 이용하면 다음과 같이 나타낼 수 있다.

$$\frac{d\exp[\theta m_1(t)]}{dt} = \theta \lambda_0$$

이 식을 풀면

$$\exp[\theta m_1(t)] = \theta \lambda_0 t + C \tag{2.7}$$

이 되고(단, C는 적분상수) 초기조건 $m(0) = 0$을 이용하면 $C = 1$이 된다.

따라서 식 (2.7)을 $m_1(t)$에 관한 식으로 유도하면 평균 값 함수가 되고 이 식을 다시 t에 관하여 미분하면 강도함수가 된다. 그 결과 식은 각각 다음과 같다.

$$m_1(t) = \frac{1}{\theta} \ln(\lambda_0 \theta t + 1) \tag{2.8}$$

$$\lambda_1(t) = \frac{\lambda_0}{\lambda_0 \theta t + 1} \tag{2.9}$$

을 얻는다. 여기서 $\beta_0 = \frac{1}{\theta}$ 와 $\beta_1 = \lambda_0 \theta$로 놓으면, 식 (2.8)

과 식 (2.9)는

$$m_1(t) = \beta_0 \ln(1 + \beta_1 t) \tag{2.10}$$

$$\lambda_1(t) = \frac{\beta_0 \beta_1}{1 + \beta_1 t} \tag{2.11}$$

와 같이 다시 표현 할 수 있다. 식 (2.8)에 의해 $m(\infty) = \infty$ 인 특성을 이용하면 이 모델은 소프트웨어 내에 잠재하는 에러수가 무한인 소프트웨어 발생 현상을 표현할 수 있는 NHPP 모형으로 나타낼 수 있다[15].

그리고 테스트 시점 t에서 소프트웨어 고장이 일어난다고 하는 가정하에서 신뢰구간 $(t, t+x]$(단, x는 임무 시간(Mission time) 사이에 소프트웨어의 에러가 일어나지 않을 소프트웨어 신뢰도(Software reliability)는 다음과 같이 알려져 있다[4, 11].

$$\begin{aligned}\hat{r}_1(x \mid t) &= \exp[-\{m(t+x) - m(t)\}] \\ &= \exp[-\beta_0 \ln(1+\beta_1(t+x) + \beta_0 \ln(1+\beta_1 t)] \\ &\quad (t \geq 0,\ x \geq 0)\end{aligned} \tag{2.12}$$

2.2 Crow(와이블) 모형

위 모형에 대한 발생확률인 강도함수는 다음과 같이 알려져 있다.

$$\lambda_2(t) = \alpha_1 \beta_2 t^{\beta_2 - 1} \quad (\text{단}, \alpha_1 > 0,\ \beta_2 > 0) \tag{2.13}$$

평균 값 함수 $m(t) = \int_0^t \lambda(x)dx$을 이용하면 다음과 같은 식으로 평균 값 함수를 나타낼 수 있다.

$$m_2(t) = \alpha_1 t^{\beta_2} \tag{2.14}$$

따라서, 소프트웨어 신뢰도 다음과 같이 나타낼 수 있다.

$$\begin{aligned}\hat{r}_2(x \mid t) &= \exp[-\mathrm{m}(t+x) + m(t)\}] \\ &= \exp[-\alpha_1 (t+x)^{\beta_2} + \alpha_1 t^{\beta_2}]\ (t \geq 0,\ x \geq 0)\end{aligned} \tag{2.15}$$

2.3 Rayleigh 모형

Glick[12]은 기록값(Record values) 열(Sequence)은 무한(Infinite)함을 제시하였다. 그러므로 $t \to \infty$ 함에 따라 고장의수가 무한하게 될 때 고장시간을 모형화 할 수 있다.

*NHPP*에 기초한 모형과 관련하여 Dwass와 Resnick[12]는 다음과 같은 정리를 제시하였다.

F(분포함수)가 실수영역 R^+에서 연속일때 구간 $(0, t]$에서 일어나는 레코드 값은 구간 $(0, t]$에서 *NHPP*의 점과정(Point process)이 되고 평균 값은 다음과 같다.

$$m_3(t) = -\ln(1 - F(t)). \tag{2.16}$$

따라서, *NHPP*의 강도함수 $\lambda_3(t) = m'(t) = f(t)/(1-F(t))$, 즉, F의 위험함수(Hazard function)가 되는 것이다. 예를 들어, 지수, 파레토(Pareto), 와이블(Weibull)를 고려하면, $f(t)/(1-F(t))$는 각 각 β, $\alpha/(\beta+t)$, $\beta\alpha t^{\alpha-1}$이 된다. 그러므로 점과정(Point processes)이 지수분포의 경우는 동질적인 포아송 과정(Homogeneous Poisson process ; HPP)이고 Musa-Okumoto, Weibull(Duane)인 경우에는 비동질적인 포아송 과정(Nonhomogeneous Poisson process ; NHPP)이 된다[11, 13].

식 (2.16)과 $\lambda(t) = m'(t) = f(t)/(1-F(t))$을 관련하면 우도함수는 다음과 같이 나타낼 수 있다. 즉, 이 식은 다음 절 식 (3.1)과 같은 형태가 된다.

$$\begin{aligned}L_{NHPP}(\beta \mid D_t) = &[\Pi_{i=1}^n f(x_i \mid \beta)/(1-F(x_i \mid \beta))] \\ &\cdot (1 - F(t \mid \beta))\end{aligned} \tag{2.17}$$

고장절단 모형에 있어서는 식 (2.17)을 이용하여 t 대신에 x_n으로 대치하면 우도함수 $L_{NHPP}(\beta \mid D_{x_n})$을 이용할 수 있다. 본 논문에서는 Rayleigh 분포 모형을 이용한 모형을 적용하고자 한다. 따라서 발생확률인 강도함수는 다음과 같이 알려져 있다[14].

$$\lambda_3(t) = m_3'(t) = f(t)/(1-F(t)) = 2\beta t \tag{2.18}$$

그리고, 평균 값 함수는 다음과 같이 나타낼 수 있다.

$$m_3(t) = -\ln(1-F(t)) = \beta_3 t^2 \tag{2.19}$$

따라서, 소프트웨어 신뢰도 다음과 같이 나타낼 수 있다.

$$\begin{aligned}\hat{r}_3(x \mid t) &= \exp[-\mathrm{m}(t+x) + m(t)\}] \\ &= \exp[-\beta_3 (t+x)^2 + \beta_3 (t)^2]\ (t \geq 0,\ x \geq 0)\end{aligned} \tag{2.20}$$

3. 모형에 대한 베이지안 모수 추정

관측시간 $(0, t]$사이에서 n번째까지 고장시점이 관찰된 고장 절단 모형일 경우($x_n = t$)를 $x_i (i = 1, 2, \cdots, n)$로 나타내면 자료 집합 D_t은 $\{x_1, x_2 \cdots x_n; t\}$으로 구성되고 x_1에서부터 x_n까지의 자료가 주어졌을 때 NHPP의 우도함수는 다음과 같이 됨이 알려져 있다[13, 14].

$$L_{NHPP} = [\prod_{i=1}^n \lambda(x_i)] \exp[-m(x_n)] \tag{3.1}$$

단, $t = x_n$(최종 고장 시점)이면 시간절단 모형(Time truncated model)이 된다. 본 논문에서는 이러한 시간절단 모형을 이용하고자 한다.

3.1 대수형 포아송 모형

대수형 포아송 실행 모형에 대한 평균 값 함수와 강도함수를 이용하면 다음과 같은 우도함수를 나타낼 수 있다.

$$L_{NHPP}(\beta_0, \beta_1 \mid D_t) = \left[\prod_{i=1}^{n} \left(\frac{\beta_0 \beta_1}{1+\beta_1 x_i} \right) \right] \exp[-\beta_0 \ln(1+\beta_1 x_n)] \quad (3.2)$$

(단, $\beta_0 = \frac{1}{\theta}$와 $\beta_1 = \lambda_0 \theta$)

이 모형에 대한 사후밀도 함수를 관찰하기 위해서는 베이즈 정리을 이용하면 다음과 같이 나타낼 수 있다.

$$f_{NHPP}(\beta_0, \beta_1 \mid D_t) \propto \left[\prod_{i=1}^{n} \left(\frac{\beta_0 \beta_1}{1+\beta_1 x_i} \right) \right] \cdot \exp[-\beta_0 \ln(1+\beta_1 x_n)] \cdot \pi_0(\beta_0) \cdot \pi_0(\beta_1) \quad (3.3)$$

(단, $\pi_0(\beta_0)$는 평균이 $\frac{a_1}{b_1}(a_1 > 0, b_1 > 0)$인 감마분포 $\Gamma(a_1, b_1)$을 따르는 β_0의 사전밀도를 나타내고 $\pi_0(\beta_1)$는 $\frac{1}{\beta_1}(0 < \beta_1 < 1)$을 따르는 β_1의 비정보사전밀도(Noninformative prior density)를 가정함)

식 (3.3)을 이용하여 모수 추정을 하기 위하여 깁스 추출법을 사용하고자 한다. 이 추출법을 이용하기 위해서는 다음과 같은 조건부 사후 분포가 필요하다.

$$\beta_0 \mid \beta_1, D_t \sim \Gamma(n+a, b+\ln(1+b_1 x_n)) \quad (3.4.1)$$

$$\beta_1 \mid \beta_0, D_t \propto \beta_1^{n-1} \frac{1}{\prod_{i=1}^{n}(1+\beta_1 x_i)} \cdot \exp[-\beta_0 \ln(1+\beta_1 x_n)] \cdot \pi_0(\beta_1) \quad (3.4.2)$$

● 깁스 알고리즘 시행 단계

위 식 (3.4.1)과 식 (3.4.2)을 이용하여 다음과 같은 단계를 이용한다. 본 논문에서는 β_0에 대한 사전분포는 비교적 넓은 범위에서 표본 발생이 이루어지도록 분산이 큰 감마분포 $\Gamma(1, 0.001)$를 주고 β_1에 대한 사전분포는 임의의 상수 예를 들면 $0 < \beta_1 < 1$를 만족하는 임의의 상수를 초기치로 준다.

① 1-1 단계

$\Gamma(1, 0.001)$의 분포에서 자료를 랜덤 발생($\beta_0^{(0)}$)시켜 식 (3.4.1)에 대입하여 랜덤 표본 $\beta_0^{(1)}$을 얻는다.

$$\beta_0^{(1)} \sim \Gamma(n+a, b+\ln(1+b_1 x_n))$$

② 1-2 단계 : 메트로폴리스 알고리즘

설명을 간결하게 하기 위해서 식 (3.4.2)의 오른쪽 식, 즉 목적 분포를 $f(\beta_1)$로 표시하자.

이 식에서 β_0는 (1-1) 단계에서 얻은 값을 대입하고 β_1는 임의의 상수 $0 < \beta_1 < 1$을 만족하는 값을 대입하고 추이 커널(Transitional kernel)은 거의 대칭을 이루는 $\Gamma(1, 0.001)$에서 β_1'을 랜덤 표본 발생하고 균등분포(0, 1]에서 임의의 확률변량을 w라 하면 $\log w \le \log f(\beta_1') - \log(\beta_1)$을 만족하면 β_1'을 $\beta_1^{(1)}$으로 간주되고 만족하지 않으면 $\beta_1^{(1)}$은 다시 β_1'으로 대치되면서 계속 반복된다.

③ 2 단계

1-1 단계와 1-2 단계를 시행하여 최근에 생성된 랜덤 표본의 값들로 대체하면서 충분히 큰 수 t만큼 반복한다. 이렇게 하여 얻은 표본을 ($\beta_0^{(t)}, \beta_1^{(t)}$)하자.

④ 3 단계

1-1 단계와 1-2 단계를 다시 (m-1)번 충분히 반복 적용하면 m개의 랜덤 표본

$$(\beta_{0(1)}^{(t)}, \beta_{1(1)}^{(t)}), (\beta_{0(2)}^{(t)}, \beta_{1(2)}^{(t)}), \cdots, (\beta_{0(m)}^{(t)}, \beta_{1(m)}^{(t)})$$

이 얻어진다.

⑤ 4 단계

최종적인 결과에 의해 β_0와 β_1의 추정은 다음과 같다.

$$\hat{\beta}_{0\ Gibbs} = \frac{1}{m}\sum_{l=1}^{m}\beta_{0(l)}^{(t)}, \quad \hat{\beta}_{1\ Gibbs} = \frac{1}{m}\sum_{l=1}^{m}\beta_{1(l)}^{(t)}$$

3.2 Crow(와이블) 모형

위 모형에 대한 평균 값 함수와 강도함수를 이용하면 다음과 식 (3.5)과 같은 우도함수를 나타낼 수 있다.

$$f_{NHPP}(\alpha_1, \beta_2 \mid D_t) = \left[\prod_{i=1}^{n} (\alpha_1 \beta_2 x_i^{\beta_2 - 1}) \right] \exp[-\alpha_1 x_n^{\beta_2}] \quad (3.5)$$

이 모형에 대한 사후밀도함수를 관찰하기 위해서는 베이즈 정리을 이용하면 다음과 같이 나타낼 수 있다.

$$f_{NHPP}(\alpha_1, \beta_2 \mid D_t) \propto \left[\prod_{i=1}^{n} (\alpha_1 \beta_2 x_i^{\beta_2 - 1}) \right] \exp[-\alpha_1 x_n^{\beta_2}] \cdot \pi_0(\alpha_1) \cdot \pi_0(\beta_2) \quad (3.6)$$

(단, $\pi_0(\alpha_1)$는 평균이 $\frac{a_2}{b_2}(a_2 > 0, b_2 > 0)$인 감마분포 $\Gamma(a_2, b_2)$을 따르는 α_1의 사전밀도를 나타내고 $\pi_0(\beta_2)$는 $\frac{1}{\beta_2}$ $(0 < \beta_2 > 1)$을 따르는 β_2의 비정보사전밀도(Noninformative prior density)를 가정함)

식 (3.6)을 이용하여 모수 추정을 하기 위하여 깁스 추출법을 사용하고자 한다. 이 추출법을 이용하기 위해서는 다음과 같은 조건부 사후 분포가 필요하다.

$$\alpha_1 \mid \beta_2, D_t \sim \Gamma(n+a_2, x_n^{\beta_2}+b) \quad (3.7.1)$$

$$\beta_2 \mid \alpha_1, D_t \propto \beta_2^{n-1}(\Pi_{i=1}^{n})^{\beta 2} \cdot e^{-\alpha_1 x_n^{\beta}} \cdot \pi_0(\beta_2) \quad (3.7.2)$$

● 깁스 알고리즘 시행 단계

식 (3.7.1)과 식 (3.7.2)을 이용하여 다음과 같은 단계를 이용한다. 본 논문에서는 α_1에 대한 사전분포는 비교적 넓은 범위에서 표본 발생이 이루어지도록 분산이 큰 감마분포 $\Gamma(1, 0.001)$를 주고 β_2에 대한 사전분포는 임의의 상수 예를 들면 $0 < \beta_2 < 1$를 만족하는 임의의 상수를 초기치로 준다.

① 1-1 단계

$\Gamma(1, 0.001)$의 분포에서 자료를 랜덤 발생($\alpha_1^{(0)}$)시켜 식 (3.7.1)에 대입하여 랜덤 표본 $\alpha_1^{(1)}$을 얻는다.

$$\beta_1^{(1)} \sim \Gamma(n + a_2, x_n^{\beta_2} + b)$$

② 1-2 단계 : 메트로폴리스 알고리즘

설명을 간결하게 하기 위해서 식 (3.7.2)의 오른쪽 식, 즉 목적 분포를 $f(\beta_2)$로 표시하자.

이 식에서 α_1는 1-1 단계에서 얻은 값을 대입하고 β_2는 임의의 상수 $0 < \beta_2 < 1$ 대입하고 추이 커널(Transitional kernel)은 거의 대칭을 이루는 $\Gamma(1, 0.001)$에서 β_2'을 랜덤 발생하고 균등분포(0, 1]에서 확률변량을 w라 하면 $\log w \leq \log f(\beta_2') - \log(\beta_2)$을 만족하면 β_2'을 $\beta_2^{(1)}$으로 간주되고 만족하지 않으면 $\beta_2^{(1)}$은 다시 β_2'으로 대치되면서 계속 반복된다.

③ 2 단계

1-1 단계와 1-2 단계를 시행하여 최근에 생성된 랜덤 표본의 값들로 대체하면서 충분히 큰 수 t 만큼 반복한다. 이렇게 하여 얻은 표본을 ($\alpha_1^{(t)}, \beta_2^{(t)}$)하자.

④ 3 단계

1-1 단계와 1-2 단계를 다시 (m - 1)번 충분히 반복 적용하면 m개의 랜덤 표본

$$(\alpha_{1(1)}^{(t)}, \beta_{2(1)}^{(t)}), (\alpha_{1(2)}^{(t)}, \beta_{2(2)}^{(t)}), \cdots, (\alpha_{1(m)}^{(t)}, \beta_{2(m)}^{(t)})$$

이 얻어진다.

⑤ 4 단계

최종적인 결과에 의해 α_1와 β_1의 추정은 다음과 같다.

$$\widehat{\alpha}_{1\,Gibbs} = \frac{1}{m}\sum_{l=1}^{m}\alpha_{1(l)}^{(t)},\ \widehat{\beta}_{2\,Gibbs} = \frac{1}{m}\sum_{l=1}^{m}\beta_{2(l)}^{(t)}$$

3.3 Rayleigh 모형

위 모형에 대한 평균값 함수와 강도함수를 이용하면 다음과 식 (3.8)과 같은 우도함수를 나타낼 수 있다.

$$f_{NHPP}(\beta_3 \mid D_t) = (\Pi_{i=1}^{n} 2\beta_3 x_i)\exp[-\beta x_n^2]. \quad (3.8)$$

이 모형에 대한 사후밀도 함수를 관찰하기 위해서는 베이즈 정리을 이용하면 다음과 같이 나타낼 수 있다.

$$f_{NHPP}(\beta_3 \mid D_t) = (\Pi_{i=1}^{n} 2\beta_3 x_i)\exp[-\beta x_n^2]\cdot\pi_0(\beta_3). \quad (3.9)$$

(단, $\pi_0(\beta_3)$는 평균이 $\frac{a_3}{b_3}(a_3 > 0, b_3 > 0)$인 감마분포 $\Gamma(a_3, b_3)$을 따르는 β_3의 사전밀도를 나타냄) 식 (3.8)을 이용하여 모수 추정을 하기 위하여 깁스 추출법을 사용하고자 한다. 이 추출법을 이용하기 위해서는 다음과 같은 조건부 사후 분포(Posterior distribution)가 필요하다.

$$\beta_3 \mid D_t \sim \Gamma(n + a_3, x_n^2 + b_3). \quad (3.10)$$

● 깁스 알고리즘 시행 단계

적당한 사전 분포를 사용하여 m번 적용을 하고 t번 반복을 한다. 사전분포는 분산이 비교적 큰 과산포 분포를 고려하여 $\beta_3 \sim \Gamma(1, 0.001)$을 선택 이용하였다.

① 변수 (β_3)의 초기값을 위의 사전분포에서 하나를 랜덤 추출하여 ($\beta_3^{(0)}$)를 정하고 모수추정을 위해서 식 (3.10)에 있는 조건부 밀도함수를 이용하자.

①-1 $\beta_3 = \beta_3^{(0)}$로 고정시켰을 경우에 β_3의 조건부 분포로부터 랜덤 표본을 하나 생성시키고 이를 $\beta_3^{(1)}$이라 표현하고

①-2 $\beta_3 = \beta_3^{(1)}$로 고정시켰을 경우에 β_3의 조건부 분포로부터 랜덤 표본을 하나 생성 시키고 이를 $\beta_3^{(2)}$이라 표현하자

② 단계 ①-1, 단계 ①-2 처럼 고정시키는 β_3의 값을 가장 최근에 생성된 랜덤 표본의 값으로 대체하면서 충분히 큰 수 t번 만큼 반복 수행 한다. 이런 시행을 하여 얻게 되는 최종 랜덤 표본을 ($\beta_3^{(t)}$)이라 한다.

③ ② 단계 다시 (m - 1)번 적용 시행하면 총 m개의 랜덤 표본

$$(\beta_{3(1)}^{(t)}, \cdots, \beta_{3(m)}^{(t)})$$

이 얻어진다.

④ 최종적인 결과에 의해 β_3의 추정은 아래와 같이 된다.

$$\widehat{E}_{Gibbs}(\beta_3) = \frac{1}{m}\sum_{k=1}^{m}\beta_{3(k)}^{(t)}$$

4. 모형의 선택

효율적 모형의 비교를 위해 고장번호(i)는 시점 x_i에 관찰된 실제고장의 수 ($n_i(x_i)$)이고 $\widehat{m}_i(x_i)$는 시점 x_i에서

추정된 고장의 누적수를 나타내므로 실제 고장수와 추정된 고장의수의 편차자승합을 계산해서 작은 모형이 효율적인 모형이라고 간주 할 수 있다. 즉, 다음과 같이 나타낼 수 있다[15].

$$C_{SSE} = \sum_{i=1}^{n} (n_i(x_i) - \widehat{m}_i(x_i))^2 \qquad (4.1)$$

단, n_i는 $(0, x_i]$사이에 관찰된 오류의 수이고 $\widehat{m}_i(x_i)$는 (깁스추정 알고리즘으로 구한) 평균 값 함수의 추정 값을 의미한다.

5. 수치적인 예

고장 시간 x_i에 대한 자료는 T1 자료[14]를 이용하여 모수를 추정하고 신뢰도를 구하고자 한다. 관측 자료는 최종 고장시간 $x_{15} = 296 (x_0 = 0)$이고 각 고장 간격시간에 대한 자료는 <표 1>에 요약하였고 본 연구는 이 자료를 이용하여 모수 추정과 신뢰도, 모형선택을 시행하고자 한다.

〈표 1〉 T1 자료

고장번호 (i)	고장간격시간 $x_i - x_{i-1}$	누적고장시간 (x_i)
1	10	10
2	9	19
3	13	32
4	11	43
5	15	58
6	12	70
7	18	88
8	15	103
9	22	125
10	25	150
11	19	169
12	30	199
13	32	231
14	25	256
15	40	296

본 연구에서는 사전분포는 즉, α_1, β_0, β_1, β_2, β_3는 감마분포 $\Gamma(1, 0.001)$을 선택 이용하여 3절에 제시한 깁스 알고리즘을 적용하였다. IMSL[17] 소프트웨어를 사용하여 각 깁스 열의 전반부 $t/2$번 반복을 제외하고 후반부 $t/2$번 반복만을 고려하는 기법 즉, 분산 분석표를 이용하는 Gelman & Rubin[8]이 제시한 MC(Markov Chain)방법을 이용하여 각 모형에 대한 사후(Posterior) 밀도의 결과를 <표 2>, <표 3>, <표 4>에 나타내었다. 이 표에서 수렴성을 확인하기 위해서 500, 2000(m)번 적용에 50, 70(t)번의 결과를 나타내었다. 그리고 각 모수에 대한 사후 평균 $\widehat{E}(\theta)$(단, θ는 모수)과 신용구간(C.I), 표준편차(S.D)를 나타내었다. 따라서 이 표에서 보여 주듯이 거의 유사한 값에 수렴함을 볼 수 있기 때문에 모형선택이나 신뢰도에 있어서 2000번 적용에 70번 반복한 사후 평균(추정값)을 이용하였다.

〈표 2〉 대수형 포아송 모형에 대한 사후 밀도

모 수	적용수(m)	반복수(t)	$\widehat{E}(\beta_0)$	S.D	95% CI
β_0	500	50	5.3241	3.8521	(2.0832, 8.8754)
		70	5.3245	3.8424	(2.4215, 8.8436)
	2000	50	5.4321	3.8322	(2.3435, 8.5852)
		70	5.4342	3.8320	(2.8581, 8.4422)
모 수	적용수(m)	반복수(t)	$\widehat{E}(\beta_1)$	S.D	95% CI
β_1	500	50	0.0516	0.0293	(0.0094, 0.0753)
		70	0.0527	0.0205	(0.0084, 0.0623)
	2000	50	0.0526	0.0188	(0.0071, 0.0695)
		70	0.0528	0.0147	(0.0023, 0.0393)

〈표 3〉 Crow 모형에 대한 사후 밀도

모 수	적용수(m)	반복수(t)	$\widehat{E}(\alpha_1)$	S.D	95% CI
α_1	500	50	1.2332	3.4103	(0.0833, 3.8752)
		70	1.2341	3.2894	(0.2215, 3.8434)
	2000	50	1.2323	3.2565	(0.3438, 3.5852)
		70	1.2342	3.1398	(0.8584, 3.4427)
모 수	적용수(m)	반복수(t)	$\widehat{E}(\beta_2)$	S.D	95% CI
β_2	500	50	0.4569	0.01389	(0.1944, 0.7537)
		70	0.4564	0.01371	(0.1347, 0.7123)
	2000	50	0.4522	0.01249	(0.1218, 0.7945)
		70	0.4524	0.01244	(0.1035, 0.7593)

〈표 4〉 Rayleigh 모형에 대한 사후 밀도

모 수	적용수(m)	반복수(t)	$\widehat{E}(\beta_3)$	S.D	95% CI
β_3	500	50	0.0001965	0.0039	(0.000084, 0.002753)
		70	0.0001964	0.0030	(0.000063, 0.002623)
	2000	50	0.0001952	0.0028	(0.000052, 0.002545)
		70	0.0001944	0.0024	(0.000043, 0.002393)

모형 선택에 있어서는 편차자승합(SSE)을 이용한 모형비교를 시행하였다. <표 5>에서는 대수형 포아송 실행 모형에 대한 베이즈 추정치(평균 값 함수)와 편차자승 값과 편차자승합을 계산하는 방법을 나타내었다.

유사한 방법으로 다른 모형에 대한 편차자승합을 계산한 결과는 <표 6>에 요약 되었다. 이 결과표에서 T1 자료를 적용한 결과 편차자승합을 최소로 하는 대수형 포아송 실행 모형이 효율적인 모형으로 간주할 수 있다.

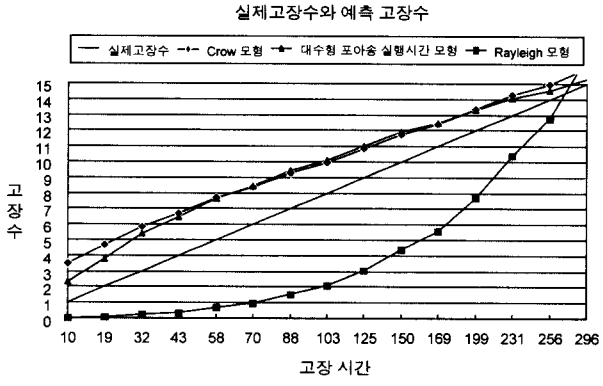
〈표 5〉 대수형 포아송 실행 모형에 대한 평균값함수의 베이즈 추정값($\widehat{m}_i(s_i)$)과 편차자승합(C_{SSE})

고장 번호 (i)	고장간격 시 간 $x_i - x_{i-1}$	누적고장 시 간 (x_i)	실 제 고장수 ($n_i(x_i)$)	평균값 함수의 추정값 ($\widehat{m}_i(x_i)$)	편차자승의 값 $(n_i(x_i)) - (\widehat{m}_i(x_i))^2$
1	10	10	1	2.3105	1.7175
2	9	19	2	5.7850	3.1862
3	13	32	3	5.3886	5.7053
4	11	43	4	6.4523	6.0138
5	15	58	5	7.6319	6.9268
6	12	70	6	8.4201	5.8570
7	18	88	7	9.4223	5.8678
8	15	103	8	10.1360	4.5627
9	22	125	9	11.0378	4.1528
10	25	150	10	11.9085	3.6422
11	19	169	11	12.4878	2.2136
12	30	199	12	13.2929	1.6715
13	32	231	13	14.0375	1.0765
14	25	256	14	14.5557	0.3088
15	40	296	15	15.2940	0.0864
C_{SSE}					52.9888

〈표 6〉 편차자승합에 의한 모형비교

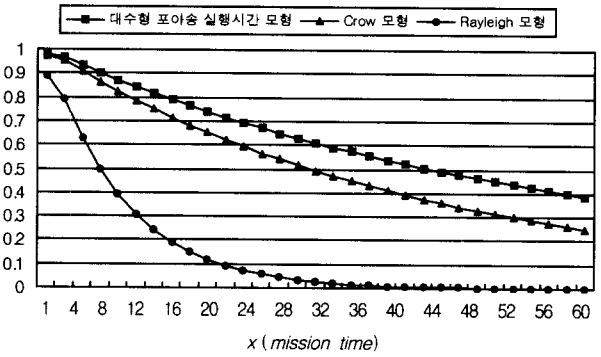
	대수형 포아송 모형	Crow 모형	Rayleigh 모형
C_{SSE}	52.9888	64.0068	262.7969

(그림 1)에서는 각 모형에 대한 예측 고장수를 도표로 그린 결과에서도 다른 모형에 비해 대수형 포아송 실행 모형이 실제 고장수에 가장 근접한 형태로 나타나고 있다.



(그림 1) 각 모형에 대한 예측 고장수

(그림 2)은 신뢰도 함수를 이용하여 미래의 신뢰도 즉, 실제고장수가 15이후의 신뢰도($\widehat{R}(x \mid s_{15} = 296)$)를 계산하여 비교하고자 한다. 이 그림에서 시간이 지날수록 모형들이 감소추세를 보이고 있고 다른 모형보다 대수형 포아송 실행 모형이 비교적 높음을 보이고 있다.



(그림 2) 각 모형에 대한 신뢰도

6. 결 론

소프트웨어의 신뢰성은 개발의 최종 단계에 있는 테스트 공정이나 실제 사용 단계에 있어서 소프트웨어 내에 존재하는 에러 수나 소프트웨어의 고장 발생시간에 의해 효과적 평가를 할 수 있는 것으로 그 평가 기술이 중요하게 된다. 소프트웨어 개발의 테스트 공정이나 실제 사용 단계에 있어서 에러 발생상황이나 소프트웨어 고장 발생현상을 수리적 모형화가 가능하다면 평가를 할 수 있다. 테스트의 개발 상황의 파악, 테스트에 의해 미 발생되었던 고장에 대한 보수비용의 예측 등 구체적인 소프트웨어 개발의 보수관리문제에도 적용 가능하다. 따라서 테스트 시간 혹은 실행시간과 발생된 고장 수나 소프트웨어 고장의 발생시간과의 관계를 소프트웨어 신뢰도 성장과정 모형에 적용 시킬 수 있다.

본 논문에서는 각 모수에 대한 사전정보가 있는 비동질 포아송 과정(NHPP)에 기초한 소프트웨어 에러 현상에 대한 베이지안 확률 모형을 고려하고 모수의 추정을 위해 잠재변수를 사용한 깁스 추정량을 이용하였다. 고장 패턴은 NHPP에 대한 강도함수와 평균값 함수로서 나타낼 수 있다. 따라서 본 논문에서는 신뢰도 측면에서 많이 사용되는 대수형 포아송 실행 모형과 Crow 모형, 그리고 Rayleigh 분포를 이용한 Rayleigh 모형을 적용하고 또, 효율적 모형을 위한 모형 선택으로서 편차자승합을 이용하여 비교하였다.

수치적인 예에서는 Musa의 T1 자료[14]를 이용하여 모수 및 신뢰도를 추정하였고 편차자승합을 이용한 모형 비교의 결과를 나열하였다.

그 결과 본 논문에서 적용한 대수형 포아송 실행 모형이 편차자승합이 다른 모형에 비해 작으므로 보다 효율적인 모형이 됨을 알 수 있었다.

신뢰도 추정을 위하여 실제고장수가 15이후의 신뢰도를 계산한 결과 임무시간(x ; Mission time)에 따른 신뢰도함수는 전체적으로 신뢰도가 비증가 추세를 보이고 있음을 알 수 있다.

따라서 대수형 포아송 실행 모형이 다른 모형에 비하여 급격히 감소하지 않은 추세를 보임으로 인해 다른 모형에 비해 신뢰도가 높음(신뢰성장)을 알 수 있다. 향후 추가적

인 분포함수를 이용한 모형 구현 및 베이지안 분석과 모수 추정에 관한 수리적 분포 이론 및 응용에 관한 연구가 기대된다.

참 고 문 헌

[1] 김희철, 이승주, " RAYLEIGH와 ERLANG 추세를 가진 혼합 고장모형에 대한 베이지안 추론에 관한 연구", 응용통계연구, 제13권 제2호, pp.505-514, 2000.

[2] 이상식, 김희철, 송영재, " NHPP에 기초한 소프트웨어 신뢰도 모형에 대한 베이지안 추론에 관한 연구 ", 정보처리학회논문지D, 제9-D권 제3호, pp.389-398, 2002.

[3] Casella, G. and George, E. I., "Explaining the Gibbs Sampler," The American Statistician, 46, pp.167-174, 1992.

[4] Chib, S ang Greenberg, E., "Understanding the Metropolis-Hastings Algorithm," The American Statistican, Vol. 49, pp.327-335, 1995.

[5] çinlar, E., "Introduction To Stochastic Process," New Jersey, Prentice-Hall, 1975.

[6] Gelfand, A. E. and Smith, A. F. M., "Sampling-Based Approaches to Calculating Marginal Densities," Journal of the American Statistical Association, 85, pp.398-409, 1990.

[7] Geman, S. and Geman, D., "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, pp.721-741, 1984.

[8] Gelman, A. E. and Rubin D., "Inference from Iterative Simulation Using Multiple Sequences," Statistical Science, 7, pp.457-472, 1992.

[9] Goel, A. L. and Okumoto, K., "Time Dependent Error Detection Rate Model for Software Reliability and Other Performance Measures," IEEE Transactions on Reliability, 28, pp.206-211, 1979.

[10] Hossain, S. A. and Dahiya, R. C., "Estimating the Parameters of a Non-homogeneous Poission-Process Model for Software Reliability," *IEEE Trans. Rel.*, Vol.R-42, No.4, pp.604-612, 1993.

[11] Kuo, L. and Yang, T. Y., "Bayesian Computation of Software Reliability," Journal of Computational and Graphical Statistics, 1995.

[12] Kuo, L. and Yang, T. Y., "Bayesian Computation for Nonhomogeneous Poisson process in Software Reliability," Journal of the American Statistical Association, 91, pp. 763-773, 1996.

[13] Lawless, J. F., "*Statistical Models and Methods for Lifetime Data*," pp.494-500, 1981.

[14] Musa, J. D., Iannino, A. and Okumoto, K., "Software Reliability : Measurement, Prediction, Application," New York, McGraw Hill, 1987.

[15] Okumoto, K., "A Statistical Method for Software Quality Control," IEEE Transactions on Software Engineering, Vol.se-11, No.12, pp.1424-1430, 1985.

[16] Tanner, M. and Wong, W., "The Calculation of Posterior Distributions by Data Augmentation," (with discussion), Journal of the American Statistical Association, 81, pp. 82-86, 1987.

[17] "*USER'S MANUAL, STAT/LIBRARY FORTRAN Subroutines for statistical analysis*," IMSL, Vol.3, pp.1050-1054, 1987.

이 상 식

e-mail : leess@songho.ac.kr

2000년 경희대학교 대학원 전자계산공학과 공학석사

2002년 경희대학교 대학원 전자계산공학과 박사과정수료

2001년~현재 송호대학 정보산업계열 전임강사

관심분야 : 소프트웨어공학, 소프트웨어신뢰성, S/W 재사용

김 희 철

e-mail : khc@songho.ac.kr

1998년 동국대학교 대학원 통계학과 이학박사

2000년~현재 송호대학 정보산업계열 조교수

관심분야 : 소프트웨어신뢰성공학, 웹프로그래밍, 전산통계

송 영 재

e-mail : yjsong@khu.ac.kr

1969년 인하대학교 전자공학과 공학사

1976년 일본 Keio 대학교 전산학과 공학석사

1980년 명지대학교 전산학과(공학박사)

1982년~1983년 미국 Maryland 대학교 객원교수

1986년~1988년 대한전자공학회 전자계산 연구회 전문위원장

1984년~1989년 전국 전산소장 협의회 부회장

1990년~1991년 일본 Keio 대학교 객원 교수

1984년~1989년 경희대학교 전자 계산소장

1993년~1995년 경희대학교 교무처장

1996년~1998년 경희대학교 공과대 학장

1999년~2000년 경희대학교 기획조정실장

2001년~현재 경희대학교 산업정보대학 원장

1976년~현재 경희대학교 컴퓨터공학과 교수

관심분야 : 소프트웨어공학, OOP/S, CASE 도구, S/W 재사용