The Korean Communications
in Statistics Vol. 10 No. 3, 2003
pp. 719-729

Bayesian Model Selection for Nonlinear Regression under Noninformative Prior

Jonghwa Na[1)] Jeongsuk Kim[2)]

Abstract

We propose a Bayesian model selection procedure for nonlinear regression models under noninformative prior. For informative prior, Na and Kim(2002) suggested the Bayesian model selection procedure through MCMC techniques. We extend this method to the case of noninformative prior. The difficulty with the use of noninformative prior is that it is typically improper and hence is defined only up to arbitrary constant. The methods, such as Intrinsic Bayes Factor(IBF) and Fractional Bayes Factor(FBF), are used as a resolution to the problem. We showed the detailed model selection procedure through the specific real data set.

Keywords : Nonlinear Regression, Noninformative prior, IBF, FBF, Importance Sampling.

1. 서론

최근에 두 개 이상의 모형에서 최적모형을 선택하는 방법으로 Kass와 Raftery(1995)에 의해 연구된 베이즈요인(Bayes factor)이 매우 효과적인 방법으로 알려지고 있다. 베이즈요인은 관측된 자료들의 집합 $\boldsymbol{D}=(x, Y)$을 가지고 모형 H_0와 모형 H_1에 대하여 다음과 같이 표현될 수 있다.

$$B_{10}=\frac{p(\boldsymbol{D}|H_1)}{p(\boldsymbol{D}|H_0)} \quad . \tag{1.1}$$

여기서 $p(\boldsymbol{D}|H_k)=\int p(\boldsymbol{D}|\theta_k, H_k)\pi(\theta_k|H_k)d\theta_k$은 각 모형에 대한 주변우도함수로써, $\pi(\theta_k|H_k)$는 모형 H_k하에서의 모수 θ_k에 대한 사전분포이고 $p(\boldsymbol{D}|\theta_k, H_k)$는 θ_k의 값이 주어졌을 때 $\boldsymbol{D}$의 분포함수이다. 베이즈요인을 이용한 사전연구로는 Carlin과 Chib(1993), Atkinson(1978), Dayal과 Dickey(1976), Lewis와 Raftery(1994), Spiegelhalter와 Smith (1982) 등이 있다. 이러한 연구에서 사용된 베이즈요인은 사전분포의 형태에 매우 민감하여 사전분포가 정상(proper) 분포인 경우에

1) Associate Professor, Dept. of Information and Statistics, Chungbuk National University, Cheongju, Korea. E-mail : cherin@chungbuk.ac.kr
2) Lecturer, Dept. of Information and Statistics, Chungbuk National University, Cheongju, Korea.

유용한 방법이므로 베이즈요인의 계산과정에서 모수에 대한 사전분포로써 정보적사전분포(informative prior)를 이용한다. 하지만 모수에 대한 사전정보가 충분하지 않은 경우에는 사전 정보를 타당하고 객관적으로 사용할 수 있는 비정보적사전분포(noninformative prior)의 사용이 요구된다.

정보적사전분포를 이용한 베이즈요인은 복잡한 유형의 비선형회귀모형에서 최적모형을 선택하는 문제에 있어서 유용하게 활용 될 수 있다. 이에 대한 대표적인 연구로 Na와 Kim(2002)은 비선형회귀모형에서 베이즈요인을 이용하여 최적모형을 선택하는 방법을 제안하였다. 이는 식 (1.1)의 계산에서 요구되는 $I=\int p(\boldsymbol{D}|\theta, H)\pi(\theta|H)d\theta$에 대한 고차원의 적분문제를 해결하기 위해 $\hat{I}=(2\pi)^{d/2}|\Sigma|^{1/2}p(\boldsymbol{D}|\hat{\theta}, H)\pi(\hat{\theta}|H)$와 같이 Laplace 근사를 실시한 후, 모수에 대한 추정치로 MCMC(Markov chain Monte Carlo) 과정을 통해 생성되는 난수에 기초한 Laplace-Metropolis 추정 과정을 이용하여 베이즈요인을 계산함으로써 최적의 비선형회귀모형을 선택하는 과정을 제시하였다.

한편, 비정보적사전분포를 이용한 베이즈요인은 비정보적사전분포가 전형적으로 비정상(improper) 분포의 형태가 되어 베이즈요인의 계산에 어려움이 따른다. 이 경우에 베이즈요인에 대한 효과적인 추정을 실시할 수 있는 방법이 바로 Berger와 Pericchi(1996)가 연구한 고유베이즈요인(Intrinsic Bayes Factor : IBF)과 O'Hagan(1995)이 연구한 부분베이즈요인(Fractional Bayes Factor : FBF) 방법이다.

본 논문에서는 비선형회귀모형에서 균일(uniform)사전분포, 제프리(Jefferys)사전분포, 참조(reference)사전분포 등의 비정보적 사전분포를 이용하여 고유베이즈요인과 부분베이즈요인을 구하고 이 결과값을 통하여 최적모형을 찾는 방법을 제시하였다. 또한 베이즈요인의 계산과정에서 요구되는 복잡한 형태의 적분계산을 해결하기 위해 주표본기법(importance sampling)을 사용하였다. 2절에서는 비정보적 사전분포하에서의 비선형회귀모형 선택에 대하여 소개하였으며, 3절에서는 실제 자료를 이용하여 비정보적 사전분포를 사용한 고유베이즈요인과 부분베이즈요인의 계산방법을 제안하고 비선형회귀모형에서 최적모형을 선택하였다.

2. 비정보적사전분포하의 비선형회귀모형 선택

비선형회귀모형의 일반적인 형태는 다음과 같다.

$$Y_i=f(x_i,\ \theta)+\varepsilon_i,\qquad \varepsilon_i\sim N(0,\ \sigma^2),\qquad i=1,2,\cdots,n. \tag{2.1}$$

여기서 모수 θ는 p 차원의 벡터이고 모수 θ와 σ에 대한 결합사후분포의 형태는 다음과 같이 주어진다.

$$p(\theta,\sigma|y)\ \propto p(y|\theta,\sigma)\pi(\theta,\sigma)=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-f(x_i,\theta))^2\right)\pi(\theta,\sigma)\quad .$$

여기서 $-\infty<\theta<\infty$ 이고 $0<\sigma<\infty$이다.

이러한 비선형회귀모형에서 최적모형을 선택하기 위해 비정보적 사전분포 $\pi_i^N(\theta_i)$를 사용하여 식(1.1)의 베이즈요인을 나타내면 다음과 같다.

$$B_{ij}^N = \frac{m_i^N(x)}{m_j^N(x)} = \frac{\int f_i(x|\theta_i)\pi_i^N(\theta_i)d\theta_i}{\int f_j(x|\theta_j)\pi_j^N(\theta_j)d\theta_j} \quad . \tag{2.2}$$

여기서 $m_i^N(x)$은 비정보적사전분포하의 주변우도함수이다.

본 논문에서는 비선형회귀모형에 대한 다음의 세가지 비정보적 사전분포를 고려하고자 한다. 먼저 식(2.1)의 모수 θ와 σ에 대한 균일사전분포는 다음과 같이 주어지며,

$$\pi(\theta,\sigma) \propto \frac{1}{\sigma}, \quad \sigma > 0, \quad \theta \in R^p \ , \tag{2.3}$$

참조사전분포와 제프리사전분포는 각각 식(2.4)와 식 (2.5)의 형태로 주어진다. (Eaves(1983).)

$$\pi(\theta,\sigma) \propto |I(\theta)|^{1/2}/\sigma \ . \tag{2.4}$$

$$\pi(\theta,\sigma) \propto |I(\theta)|^{1/2}/\sigma^{(p+1)} \ . \tag{2.5}$$

여기서 정보행렬(information matrix) $I(\theta) = \{E(D_\theta \|y-f(\theta)\|^2|\theta)\}/2 = d_\theta f(\theta)' d_\theta f(\theta)$로서 이때 D_θ는 이차의 $p\times p$ 편미분 행렬이고 $d_\theta f(\theta)$는 f의 $n\times p$ 야코비안(Jacobian) 행렬이다.

식(2.2)의 베이즈요인에서 두 사전분포는 일반적으로 비정상(improper) 사전분포의 형태 $\pi_i^N(\theta_i) = c_i g_i(\theta_i)$를 가짐으로써 다음과 같이 주어진다.

$$B_{ij}^N = \frac{m_i^N(x)}{m_j^N(x)} = \frac{c_i \int f_i(x|\theta_i)g_i(\theta_i)d\theta_i}{c_j \int f_j(x|\theta_j)g_j(\theta_j)d\theta_j} \quad .$$

여기서 $g_i(\theta_i)$는 Θ_i에서 적분값이 발산하는 함수이고 이때 결정되지 않는 임의의 상수 c_i와 c_j를 포함하게 된다. 이러한 문제를 해결할 수 있는 방법이 시험표본(training sample)을 사용하는 고유베이즈요인과 부분베이즈요인이다.

2.1 고유베이즈요인

먼저 자료를 $0 < l < n$인 $x = (x(l), x(n-l))$ 두 부분으로 나누어 놓고, $x(l)$은 $\pi_i^N(\theta_i)$를 상수 c_i를 포함하지 않는 적절한 사후분포 $\pi_i^N(\theta_i|x(l)) = f_i(x(l)|\theta_i)\pi_i^N(\theta_i)/m_i^N(x(l))$로 변환하는 시험표본으로 사용한다. 여기서 $m_i^N(x(l)) = \int f_i(x(l)|\theta_i)\pi_i^N(\theta_i)d\theta_i$ 이 되고, $f_i(x(l)|\theta_i)$은 모형 H_i에서 $X(l)$에 대한 주변분포가 된다. 위 식의 $\pi_i^N(\theta_i|x(l))$을 사전분포로 이용하고, 나머지 자료인 $x(n-l)$을 가지고 베이즈요인을 계산하게 된다.

고유베이즈요인은 $x(l)$이 주어져 있고, $\pi_i^N(\theta_i|x(l))$이 정상(proper) 분포라는 가정하에서 다음과 같이 표현된다.

$$B_{ij}(x(l)) = B_{ij}^N \cdot B_{ji}^N(x(l)) \ . \tag{2.6}$$

여기서 $B_{ji}^N(x(l)) = m_j^N(x(l))/m_i^N(x(l))$ 이 되어 상수 c_i와 c_j는 서로 상쇄되어 없어지게 된

다. 그리고 시험표본을 이용한 베이즈요인의 계산은 $m_i^N(x(l))$이 유한(finite)한 경우에 가능하므로 시험표본 $x(l)$은 모든 모형 H_i에 대해 $0 < m_i^N(x(l)) < \infty$을 만족할 때 정상성(proper)을 가진다. 또한 $x(l)$이 정상성을 가지고 $x(l)$보다 작은 부분시험표본이 더 이상 정상성을 가지지 않는다면 $x(l)$은 최소성(minimal)을 가지게 된다.

$B_{ij}(x(l))$은 임의의 최소시험표본(minimal training sampling) 선택에 따라 많은 영향을 받는다. 따라서 이러한 의존성을 제거하고 안정성을 높이기 위해서 모든 가능한 최소시험표본 집합 $\chi_T = \{x(1), x(2), \cdots, x(L)\}$에 대하여 식(2.7)과 같이 산술평균 고유베이즈요인(Arithmetic Intrinsic Bayes Factor : AIBF), 기하평균 고유베이즈요인(Geometric Intrinsic Bayes Factor : GIBF), 표본의 수가 적은 경우에 유용한 중위수 고유베이즈요인(Median Intrinsic Bayes Factor; MIBF) 방법을 사용할 수 있다.

$$
\begin{aligned}
B_{ij}^{AIBF} &= \frac{1}{L}\sum_{l=1}^{L} B_{ij}(x(l)) = B_{ij}^N \cdot \frac{1}{L}\sum_{l=1}^{L} B_{ji}^N(x(l)) \ , \\
B_{ij}^{GIBF} &= \left\{ \prod_{l=1}^{L} B_{ij}(x(l)) \right\}^{1/L} = B_{ij}^N \cdot \left\{ \prod_{l=1}^{L} B_{ji}^N(x(l)) \right\}^{1/L} \ , \\
B_{ij}^{MIBF} &= B_{ij}^N \cdot Med\{B_{ji}^N(x(l))\} \ .
\end{aligned} \tag{2.7}
$$

2.2 부분베이즈요인

비정보적(noninformative) 사전분포가 비정상 사전분포의 형태를 가진 경우에 시험표본을 이용하여 베이즈요인을 구하는 또 다른 방법으로 O'Hagan(1995)이 제안한 부분베이즈요인이 있다. 이 방법은 임의의 특정한 시험표본 $x(l)$을 선택해야 되는 문제 또는 고유베이즈요인 방법에서 모든 가능한 최소시험표본 집합을 고려해야만 하는 문제 등을 해결하고 좀더 간단한 형태의 베이즈요인을 계산할 수 있다.

먼저 시험표본이 차지하는 비율을 $b = l/n$이라 놓고 n과 l이 매우 크다고 하다면 $f_i(x(l)|\theta_i)$은 $f_i(x|\theta_i)^b$에 근사하게 되어 다음과 같이 표현된다.

$$
f_i(x(l)|\theta_i)^{\frac{n}{l}} \approx f_i(x|\theta_i) \quad \text{그리고} \quad f_i(x(l)|\theta_i) \approx f_i(x|\theta_i)^b \ .
$$

따라서 O'Hagan(1995)은 식(2.6)을 다음과 같은 부분베이즈요인으로 나타내었다.

$$
B_{ij}^F(x) = \frac{B_{ij}^N(x)}{B_{ij}^b(x)} \ . \tag{2.8}
$$

여기서 $B_{ij}^b(x) = \dfrac{m_i^b(x)}{m_j^b(x)} = \dfrac{\int f_i(x|\theta_i)^b \pi_i^N(\theta_i) d\theta_i}{\int f_j(x|\theta_j)^b \pi_j^N(\theta_j) d\theta_j}$ 이다.

만약 $\pi_i^N(\theta_i)$가 비정상 사전분포의 형태를 가진다고 해도 결정되지 않는 상수 c_i와 c_j는 식

(2.2)를 이용하여 식(2.8)에서는 서로 상쇄되어 없어지게 된다. 부분베이즈요인을 정의하는데 있어서 중요한 점은 b의 선택 방법에 대한 문제이다. O'Hagan(1995)은 b를 선택하는데 있어서 다음 세 가지 방법을 제시하였다. 여기서 l은 최소시험표본의 크기이다.

① $b=l/n$, 로버스트성을 고려하지 않을 때

② $b=\max\{l, \sqrt{n}\}/n$, 로버스트성을 고려할 때

③ $b=\max\{l, \log n\}/n$, ①과 ②를 적당히 고려할 때

3. 실제 자료에 대한 비선형회귀모형의 선택

3.1 BOD 자료와 모형소개

이 장에서는 Marske(1967)가 어떤 하천에 대하여 시간의 변화에 따른 생화학적 산소요구량(BOD)을 측정한 표3.1의 자료를 이용하여 균일사전분포, 제프리사전분포, 참조사전분포 이 세 가지의 비정보적 사전분포의 형태에 따라 고유베이즈요인과 부분베이즈요인을 계산하고 최적모형을 선택하였다.

표 3.1 : 시간에 따른 생화학적 산소요구량(BOD) 자료

시간(days)	생화학적 산소요구량(mg/l)	시간(days)	생화학적 산소요구량(mg/l)
1	8.3	4	16.0
2	10.3	5	15.6
3	19.0	7	19.8

Bates와 Watts(1988)가 제안한 비내포모형(nonnested model)의 비선형회귀모형 함수 $f(x, \theta)$를 이용하여 베이지안 방법으로 최적모형을 선택한다.

[모형1] $f(x, \theta)=\theta_1(1-e^{-\theta_2 x})$

[모형2] $f(x, \theta)=\dfrac{\theta_1 x}{\theta_2+x}$

3.2 비정보적사전분포하의 베이즈요인

본 논문에서는 세 가지 비정보적사전분포의 형태에 따라 산술평균을 이용한 고유베이즈요인과 중위수를 이용한 고유베이즈요인의 값을 계산하였고, O'Hagan이 제시한 b를 선택하는 세 가지 방법에 따라 부분베이즈요인을 구해보았다. 베이즈요인을 계산할 때 주변우도함수 $m_i^N(x)$의 복잡한 적분계산을 위하여 주표본기법(importance sampling)을 사용한다. 이때 주표본함수는 이변량표준정규분포를 사용하여 이로부터 10000개의 θ들을 추출하였다. 모든 가능한 최소시험표본 집합 $y_T=\{y(1), y(2), \cdots, y(L)\}$의 최소시험표본 $y(l)$, $l=1, 2, \cdots, L$은 각각 크기가 2인 표본

$(y_i, y_j), i \neq j$ 을 의미하며 $L=\binom{6}{2}=15$이다. 부분베이즈요인의 계산에서는 최소시험표본 l의 값이 2가 되므로 $b=l/n$ 과 $b=\max\{l, \log n\}/n$는 같은 값을 가지게 된다.

3.2.1 균일사전분포를 이용한 베이즈요인

이 절은 3.1절의 두 모형의 모수 θ와 σ에 대하여 식(2.3)의 균일사전분포를 이용하여 고유베이즈요인과 부분베이즈요인을 계산하는 방법이다. 먼저 식(2.7)의 고유베이즈요인을 계산하기 위해서 사용되는 B_{12}^N와 $B_{21}^N(y(l))$를 계산하면 다음과 같다.

$$B_{12}^N=\frac{m_1^N(y)}{m_2^N(y)}=\frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_0^{\infty}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta_1(1-e^{-\theta_2 x_i}))^2\right)\frac{1}{\sigma}\,d\sigma d\theta_1 d\theta_2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_0^{\infty}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right)\frac{1}{\sigma}\,d\sigma d\theta_1 d\theta_2}$$

$$=\frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i=1}^{n}(y_i-\theta_1(1-e^{-\theta_2 x_i}))^2\right)^{-\frac{n}{2}}d\theta_1 d\theta_2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i=1}^{n}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right)^{-\frac{n}{2}}d\theta_1 d\theta_2}. \tag{3.1}$$

$$B_{21}^N(y(l))=\frac{m_2^N(y(l))}{m_1^N(y(l))}=\frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i\in y(l)}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right)^{-1}d\theta_1 d\theta_2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i\in y(l)}(y_i-\theta_1(1-e^{-\theta_2 x_i}))^2\right)^{-1}d\theta_1 d\theta_2}.$$

식(2.8)의 부분베이즈요인을 계산하기 위해서 B_{12}^N은 식(3.1)과 같고 $B_{12}^b(y)$는 다음과 같이 계산된다.

$$B_{12}^b(y)=\frac{m_1^b(y)}{m_2^b(y)}$$

$$=\frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_0^{\infty}\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta_1(1-e^{-\theta_2 x_i}))^2\right)\right)^b\frac{1}{\sigma}\,d\sigma d\theta_1 d\theta_2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_0^{\infty}\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right)\right)^b\frac{1}{\sigma}\,d\sigma d\theta_1 d\theta_2}$$

$$=\frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i=1}^{n}(y_i-\theta_1(1-e^{-\theta_2 x_i}))^2\right)^{-\frac{nb}{2}}d\theta_1 d\theta_2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i=1}^{n}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right)^{-\frac{nb}{2}}d\theta_1 d\theta_2}. \tag{3.2}$$

3.2.2 참조사전분포를 이용한 베이즈요인

이 절은 3.1절의 두 모형의 모수 θ와 σ에 대하여 식(2.4)의 참조사전분포를 이용하여 고유베이즈요인과 부분베이즈요인을 계산하는 방법이다. 먼저 고유베이즈요인을 계산하기 위해 두 모형에 대한 사전분포 $\pi_1(\theta,\sigma)$와 $\pi_2(\theta,\sigma)$는 다음의 식(3.3)과 같이 표현된다.

$$\pi_1(\theta,\sigma) = \frac{1}{\sigma}\left(\sum_{i=1}^{n}(1-e^{-\theta_2 x_i})^2\sum_{i=1}^{n}(\theta_1 x_i e^{-\theta_2 x_i})^2-\left(\sum_{i=1}^{n}\theta_1 x_i e^{-\theta_2 x_i}(1-e^{-\theta_2 x_i})\right)^2\right)^{1/2}.$$

$$\pi_2(\theta,\sigma) = \frac{1}{\sigma}\left(\sum_{i=1}^{n}\frac{x_i^2}{(\theta_2+x_i)^2}\sum_{i=1}^{n}\frac{\theta_1^2 x_i^2}{(\theta_2+x_i)^4}-\left(\sum_{i=1}^{n}\frac{-\theta_1 x_i^2}{(\theta_2+x_i)^3}\right)^2\right)^{1/2}. \tag{3.3}$$

식(3.3)의 사전분포를 이용하여 B_{12}^N과 $B_{21}^N(y(l))$은 다음의 식(3.4)와 같이 계산된다.

$$B_{12}^N = \frac{m_1^N(y)}{m_2^N(y)}$$

$$= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_0^{\infty}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta_1(1-e^{-\theta_2 x_i}))^2\right)\times\pi_1(\theta,\sigma)\,d\sigma\,d\theta_1 d\theta_2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_0^{\infty}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right)\times\pi_2(\theta,\sigma)\,d\sigma\,d\theta_1 d\theta_2}$$

$$= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i=1}^{n}(y_i-\theta_1(1-e^{-\theta_2 x_i}))^2\right)^{-\frac{n}{2}}\cdot p_1(\theta_1,\theta_2)\,d\theta_1 d\theta_2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i=1}^{n}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right)^{-\frac{n}{2}}\cdot p_2(\theta_1,\theta_2)\,d\theta_1 d\theta_2}. \tag{3.4}$$

$$B_{21}^N(y(l)) = \frac{m_2^N(y(l))}{m_1^N(y(l))} = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i\in y(l)}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right)^{-1}\cdot p_1(\theta_1,\theta_2)\,d\theta_1 d\theta_2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i\in y(l)}(y_i-\theta_1(1-e^{-\theta_2 x_i}))^2\right)^{-1}\cdot p_2(\theta_1,\theta_2)\,d\theta_1 d\theta_2}.$$

여기서

$$p_1(\theta_1,\theta_2) = \left(\sum_{i=1}^{n}(1-e^{-\theta_2 x_i})^2\sum_{i=1}^{n}(\theta_1 x_i e^{-\theta_2 x_i})^2-\left(\sum_{i=1}^{n}\theta_1 x_i e^{-\theta_2 x_i}(1-e^{-\theta_2 x_i})\right)^2\right)^{1/2}$$

$$p_2(\theta_1,\theta_2) = \left(\sum_{i=1}^{n}\frac{x_i^2}{(\theta_2+x_i)^2}\sum_{i=1}^{n}\frac{\theta_1^2 x_i^2}{(\theta_2+x_i)^4}-\left(\sum_{i=1}^{n}\frac{-\theta_1 x_i^2}{(\theta_2+x_i)^3}\right)^2\right)^{1/2} \tag{3.5}$$

이다.

부분베이즈요인을 계산하기 위해 두 모형의 사전분포 $\pi_1(\theta,\sigma)$와 $\pi_2(\theta,\sigma)$는 식(3.3)과 같고, B_{12}^N는 식(3.4)와 같이 표현되며 $B_{12}^b(y)$는 다음의 식(3.6)과 같이 계산된다.

$$B_{12}^{b}(y) = \frac{m_1^b(y)}{m_2^b(y)}$$
$$= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty}\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta_1(1-e^{-\theta_2 x_i}))^2\right)\right)^b \times \pi_1(\theta,\sigma)\, d\sigma d\theta_1 d\theta_2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty}\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right)\right)^b \times \pi_2(\theta,\sigma)\, d\sigma d\theta_1 d\theta_2}$$
$$= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i=1}^{n}(y_i-\theta_1(1-e^{-\theta_2 x_i}))^2\right)^{-\frac{nb}{2}} \cdot p_1(\theta_1,\theta_2)\, d\theta_1 d\theta_2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i=1}^{n}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right)^{-\frac{nb}{2}} \cdot p_2(\theta_1,\theta_2)\, d\theta_1 d\theta_2}. \tag{3.6}$$

여기서 $p_1(\theta_1,\theta_2)$과 $p_2(\theta_1,\theta_2)$는 식(3.5)와 동일하다.

3.2.3 제프리사전분포를 이용한 베이즈요인

이 절은 3.1절의 두 모형의 모수 θ와 σ에 대하여 식(2.5)의 제프리사전분포를 이용하여 고유베이즈요인과 부분베이즈요인을 계산하는 방법이다. 고유베이즈요인을 계산하기 위해 두 모형에 대한 사전분포 $\pi_1(\theta,\sigma)$와 $\pi_2(\theta,\sigma)$는 식(3.7)과 같이 표현된다.

$$\pi_1(\theta,\sigma) = \frac{1}{\sigma^3}\left(\sum_{i=1}^{n}(1-e^{-\theta_2 x_i})^2\sum_{i=1}^{n}(\theta_1 x_i e^{-\theta_2 x_i})^2 - \left(\sum_{i=1}^{n}\theta_1 x_i e^{-\theta_2 x_i}(1-e^{-\theta_2 x_i})\right)^2\right)^{1/2}.$$
$$\pi_2(\theta,\sigma) = \frac{1}{\sigma^3}\left(\sum_{i=1}^{n}\frac{x_i^2}{(\theta_2+x_i)^2}\sum_{i=1}^{n}\frac{\theta_1^2 x_i^2}{(\theta_2+x_i)^4} - \left(\sum_{i=1}^{n}\frac{-\theta_1 x_i^2}{(\theta_2+x_i)^3}\right)^2\right)^{1/2}. \tag{3.7}$$

위 사전분포를 사용하여 B_{12}^N과 $B_{21}^N(y(l))$을 계산하면 식(3.8)과 같다.

$$B_{12}^{N} = \frac{m_1^N(y)}{m_2^N(y)}$$
$$= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta_1(1-e^{-\theta_2 x_i}))^2\right)\times \pi_1(\theta,\sigma)\, d\sigma\, d\theta_1 d\theta_2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right)\times \pi_2(\theta,\sigma) d\sigma d\theta_1 d\theta_2}$$
$$= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i=1}^{n}(y_i-\theta_1(1-e^{-\theta_2 x_i}))^2\right)^{-\frac{n+2}{2}} \cdot p_1(\theta_1,\theta_2)\, d\theta_1 d\theta_2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i=1}^{n}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right)^{-\frac{n+2}{2}} \cdot p_2(\theta_1,\theta_2)\, d\theta_1 d\theta_2}. \tag{3.8}$$

$$B_{21}^{N}(y(l)) = \frac{m_2^N(y(l))}{m_1^N(y(l))} = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i\in y(l)}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right)^{-2}\cdot p_1(\theta_1,\theta_2)\,d\theta_1 d\theta_2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i\in y(l)}(y_i-\theta_1(1-e^{-\theta_2 x_i}))^2\right)^{-2}\cdot p_2(\theta_1,\theta_2)\,d\theta_1 d\theta_2}.$$

여기서

$$p_1(\theta_1,\theta_2) = \left(\sum_{i=1}^{n}(1-e^{-\theta_2 x_i})^2\sum_{i=1}^{n}(\theta_1 x_i e^{-\theta_2 x_i})^2-\left(\sum_{i=1}^{n}\theta_1 x_i e^{-\theta_2 x_i}(1-e^{-\theta_2 x_i})\right)^2\right)^{1/2}$$

$$p_2(\theta_1,\theta_2) = \left(\sum_{i=1}^{n}\frac{x_i^2}{(\theta_2+x_i)^2}\sum_{i=1}^{n}\frac{\theta_1^2 x_i^2}{(\theta_2+x_i)^4}-\left(\sum_{i=1}^{n}\frac{-\theta_1 x_i^2}{(\theta_2+x_i)^3}\right)^2\right)^{1/2} \tag{3.9}$$

이다.

부분베이즈요인을 계산하기 위해 두 모형의 사전분포 $\pi_1(\theta,\sigma)$와 $\pi_2(\theta,\sigma)$는 식(3.7)과 같고 B_{12}^N는 식(3.8)로 표현되며 $B_{12}^b(y)$는 다음 식(3.10)과 같이 계산된다.

$$\begin{aligned}
B_{12}^b(y) &= \frac{m_1^b(y)}{m_2^b(y)} \\
&= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_0^{\infty}\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta_1(1-e^{-\theta_2 x_i}))^2\right)\right)^b\times\pi_1(\theta,\sigma)\,d\sigma d\theta_1 d\theta_2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_0^{\infty}\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right)\right)^b\times\pi_2(\theta,\sigma)\,d\sigma d\theta_1 d\theta_2} \\
&= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i=1}^{n}(y_i-\theta_1(1-e^{-\theta_2 x_i}))^2\right)^{-\frac{nb+2}{2}}\cdot p_1(\theta_1,\theta_2)\,d\theta_1 d\theta_2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\sum_{i=1}^{n}\left(y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)^2\right)^{-\frac{nb+2}{2}}\cdot p_2(\theta_1,\theta_2)\,d\theta_1 d\theta_2}.
\end{aligned} \tag{3.10}$$

여기서 $p_1(\theta_1,\theta_2)$과 $p_2(\theta_1,\theta_2)$는 식(3.9)와 동일하게 표현된다.

3.3 베이즈요인을 통한 모형선택

3.2절에서 보여진 사전분포의 형태에 따라 고유베이즈요인과 부분베이즈요인의 계산을 수행한 결과 다음의 표3.2와 같은 결과를 얻었다. Jeffreys(1961)의 베이즈요인값에 따른 모형선택기준에 따라 이 결과를 살펴보면 베이즈요인의 값들이 모두 기준치 3보다 작은 값을 가지므로 제시되었던 [모형1]과 [모형2]가 모두 시간에 따른 생화학적 산소요구량(BOD) 자료에 적합한 최적모형으로 선택되었다. 균일사전분포를 고려한 경우에는 산술평균 고유베이즈요인, 중위수 고유베이즈요인, b값의 선택방법 세가지에 따른 부분베이즈요인 값들에서 모두 [모형1]과 [모형2]가 최적모형으로 선택되었으며, 참조사전분포와 제프리사전분포를 고려한 경우에도 산술평균 및 중위수 고유베이즈요인과 부분베이즈요인의 계산된 값들로부터 [모형1]과 [모형2]가 최적모형으로 선택되었다. 또한 비선형회귀모형의 고전적인 모형선택법은 주로 내포모형(nested model)에 대하여만 적용가능하였는데 베이즈요인을 이용한 베이지안 접근 방식의 모형선택은 비내포모형(nonnested model)에 대해서도 용이하게 적용할 수 있다는 것을 알 수 있다.

표 3.2 : 사전분포의 형태에 따른 IBF와 FBF의 계산 결과

베이즈요인 \ 사전분포		균일(uniform)	참조(reference)	제프리(Jefferys)
$B_{12}^{AIBF}(y)$		1.3056	1.8849	1.3027
$B_{12}^{MIBF}(y)$		1.0861	1.1682	1.6862
$B_{12}^{F}(y)$	$b=\frac{l}{n}$	1.0146	2.1165	1.9512
	$b=\frac{\max\{l,\sqrt{n}\}}{n}$	1.2579	1.8718	1.6803
	$b=\frac{\max\{l,\log n\}}{n}$	1.0146	2.1165	1.9512

4. 결론

본 논문에서는 비정보적 사전분포하에서의 비선형회귀모형의 선택과정을 제시하였다. 이는 정보적사전분포하에서의 Na와 Kim(2002)의 확장으로, IBF, FBF 등을 통한 베이즈요인의 효과적인 추정을 통해 수행되었으며, 이들 베이즈요인의 계산과정에 발생하는 주변우도함수의 적분계산은 주표본기법을 사용하여 해결하였다. 여러 가지 베이즈요인은 자료의 개수나 사전분포에 매우 민감하게 작용하므로 이들의 변화에 따른 민감도분석(sensitivity analysis) 및 로버스트성(robustness)에 대한 추가적인 연구가 필요하다.

참고문헌

[1] Atkinson, A. C. (1978). Posterior Probabilities for Choosing a Regression Model. *Biometrika*, 65, 39-48.

[2] Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and its Applications*, John Wiley & Sons.

[3] Berger, J. O. and Pericchi, L. R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction, *Journal of the American Statistical Association*, 91, 109-121.

[4] Carlin, B. P. and Chib, S. (1993). Bayesian Model Choice via Marcov Chain Monte Carlo. *Research Report* 93-006, University of Minnesota, Division of Biostatistics.

[5] Dayal, H. H. and Dickey, J. M. (1976). Bayes Factors for Behrens-Fisher Problems. *Sankhya*, 38, 315-328.

[6] Eaves, D. M. (1983). On Bayesian nonlinear regression with an enzyme example. *Biometrika*, 70, 2, 373-379.

[7] Jeffreys, H. (1961). *Theory of Probability*(3rd ed.), Oxford, U.K. : Oxford University Press.

[8] Kass, R. E. and Raftery, A. E. (1995). Bayes factors, *Journal of the American Statistical Association*, 90, 377-395.

[9] Lewis, S. M. and Raftery, A. E. (1994). The Laplace-Metropolis Estimator for Bayes Factors via Posterior Simulation. *Technical Report* 279, University of Washington, Dept. of Statistics.

[10] Marske, Donald. (1967). *Biochemical Oxygen Demand Data Interpretation Using Sum of Squares Surface,* M.S. Thesis, University of Wisconsin-Madison.

[11] Na, J. H. and Kim, J. S. (2002). Bayesian Model Selection and Diagnostics for Nonlinear Regression Model via MCMC. *The Korean Journal of Applied Statistics,* 15, 1, 139-151.

[12] O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparisons, *Journal of the Royal Statistical Society*, Ser. B, 57, 99-138.

[13] Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes Factors for Linear and Log-Linear Models with Vague Prior Information. *Journal of the Royal Statistical Society*. Ser. B, 44, 377-387.

[2003년 6월 접수, 2003년 9월 채택]