응용통계연구 제 15권 1호, 2002년, pp. 139-151

베이지안 비선형회귀모형의 선택과 진단 *

나종화 [1] 김정숙 [2]

요 약

본 논문에서는 베이지안 기법을 이용한 비선형회귀모형의 선택법을 제안하였다. 베이즈요인에 기초한 이 방법은 주로 대표본의 경우에 이용되는 고전적 모형선택법에 비해 사전정보를 이용하는 측면과 비내포모형 및 소표본의 경우에 대해서도 효과적으로 사용될 수 있다는 장점을 가진다. 본 논문에서는 정보적 사전분포를 고려하였으며, 베이즈요인의 추정방법으로 Laplace-Metropolis 추정법을 제안하였다. 또한 MCMC 과정을 통해 추정된 모수의 수렴진단에 대해서도 고려하였다. 실제자료에 대한 최적의 모형선택 및 진단과정을 구체적으로 제시하였다.

주요용어: MCMC, 비선형회귀모형, 베이즈요인, Laplace-Metropolis추정량, 수렴진단.

1. 서론

공학이나 의학 등의 분야에서 나오는 대부분의 통계적인 실험 자료에 대한 최적의 비선형회귀모형(nonlinear regression model)의 선택은 대단히 중요하다. 그러나 복잡한 유형의 비선형회귀모형에서 최적모형을 선택하기란 쉬운 일이 아니며 이에 대한 연구도 미비한 실정이다.

비선형회귀모형의 일반적인 형태는 다음과 같다.

$$Y_i = f(x_i, \theta) + \epsilon_i, \quad i = 1, 2, \cdots, n. \tag{1.1}$$

여기서 θ는 P차원벡터이고 $f(x, \theta)$는 일차원 독립변수 x와 종속변수 Y의 관계를 나타내는 θ에 대한 적절한 비선형함수가 된다. 또한 $E(\epsilon_i) = 0$이며 $Var(\epsilon_i) = \sigma^2$이다. 위 식(1.1)과 같은 비선형회귀모형의 고전적인 모형선택법으로는 네이만-피어슨(Neyman-Pearson) 정리를 이용하는 방법과 우도비검정(likelihood ratio test)을 이용하는 방법이 주로 사용되어 지고 있다. 하지만 이러한 방법에 기초한 모형선택은 내포모형(nested model)에 대해서만 이용할 수 있고, 근사적 χ^2값에 기초하여 모형선택을 수행하기 때문에 대표본인 경우에는 적합하나 소표본인 경우에는 적용하기가 어렵다는 단점을 가지고 있다. 그러나 베이즈요인을 이용한 베이지안 접근방식의 모형선택은 내포모형 또는 비내포모형(nonnested model) 모

* 본 연구는 한국과학재단 목적기초연구(2001-1-10400-017-1) 지원으로 수행되었음.

1) (361-763) 충북 청주시 흥덕구 개신동 산 48, 충북대학교 통계학과 부교수
E-mail : cherin@cbucc.chungbuk.ac.kr
2) (361-763) 충북 청주시 흥덕구 개신동 산 48, 충북대학교 통계학과 박사과정수료
E-mail : chastity@trut.chungbuk.ac.kr

두에 대해서도 활용가능 하다는 장점을 가지고 있고, 또한 자료의 수가 적은 경우에도 적합하게 사용할 수가 있다. 최근에는 소표본의 자료이거나 복잡한 유형의 모형을 분석하는데 있어서 모수에 대한 사전정보(prior information)와 자료의 정보를 모두 이용하는 베이지안 기법이 널리 이용되고 있다. 특히 Kass와 Raftery(1995)에 의해 연구된 베이즈요인(Bayes factor)은 두 개 이상의 모형에서 최적모형을 선택하는데 있어서 매우 효과적인 방법으로 제시되고 있다.

베이즈요인의 계산과정에서 요구되는 고차원이거나 복잡한 형태의 함수는 적분이 어렵거나 불가능하기 때문에 이러한 어려움을 개선하기 위한 방법으로 Tierney와 Kadane(1986)에 의해 제안된Laplace근사방법 (Laplace approximation method), 주표본방법(importance sampling method), 마코프연쇄 몬테카를로 (Makov chain Monte Carlo : MCMC)방법 등을 이용하여 해결한다. 그 중 MCMC방법은 사후분포로 부터 모수 θ들의 추출을 통한 베이지안 추론을 실시하는 최근 가장 주목 받고 있는 방법중의 하나이다.

본 논문에서는 비선형회귀모형에서 MCMC방법을 이용하여 베이즈요인을 계산하고 이 결과값을 통하여 최적모형을 찾는 방법과 Gelman과 Rubin(1992) 그리고 Raftery와 Lewis(1992) 방법을 이용하여 사후분포로부터 추출된 모수들에 대한 수렴진단 방법을 제시하였다. 2절에서는 베이즈요인의 정의와 이의 계산에 요구되는 Metropolis-Hastings 알고리즘을 소개하였으며, 3절에서는 베이즈요인을 계산하기 위해서 Laplace 근사법에 기초한 Laplace-Metropolis 추정법을 제안하고 이를 이용하여 비선형회귀모형에서 최적모형을 선택하고 진단하는 방법을 제시하였다. 4절에서는 실제자료와 모의자료를 이용하여 비선형회귀모형에서 최적모형을 선택하고 진단하였다.

2. 베이즈요인의추정

2.1. 베이즈요인

먼저 n개의 관측된 자료들의 집합을 $\mathbf{D} = (x, Y)$라 하자. 모형 H_0와 모형H_1에 대한 각각의 확률밀도함수를 $p(\mathbf{D}|H_0)$과 $p(\mathbf{D}|H_1)$라고 하면 이 두 모형의 사후분포(posterior distribution) $p(H_0|\mathbf{D})$와 $p(H_1|\mathbf{D})$는 다음과 같이 표현된다.

$$p(H_k|\mathbf{D}) = p(\mathbf{D}|H_k) \cdot p(H_k), \quad k = 0, 1. \tag{2.1}$$

여기서 $p(H_0)$과 $p(H_1)$는 사전분포(prior distribution)가 된다. 베이즈정리를 이용하여 식(2.1)을 사후분포의 오즈비(odds ratio)로 변환하면 다음의 관계식을 가진다.

$$\frac{p(H_1|\mathbf{D})}{p(H_0|\mathbf{D})} = \frac{p(\mathbf{D}|H_1)}{p(\mathbf{D}|H_0)} \cdot \frac{p(H_1)}{p(H_0)} \quad . \tag{2.2}$$

이때 식(2.2)의 우변에 표현된 우도함수의 오즈를 베이즈요인(Bayes Factor)이라고 하며 다음과 같은 관계로 표현된다.

$$B_{10} = \frac{p(\mathbf{D}|H_1)}{p(\mathbf{D}|H_0)} \quad . \tag{2.3}$$

이는 모형 H_1을 모형 H_0와 비교하여 모형 H_1이 선택되기 위한 상대적 가중치를 나타내는 값이 된다.

식(2.3)에 표현된 베이즈요인을 구하기 위해서는 다음과 같은 형태의 적분에 대한 계산이 요구된다.

$$p(\mathbf{D}|H_k) = \int p(\mathbf{D}|\theta_k, H_k)\pi(\theta_k|H_k)d\theta_k \quad . \tag{2.4}$$

여기서 θ_k는 모형H_k하에서 모수이고, $\pi(\theta_k|H_k)$는 모형H_k하에서의 모수θ_k에 대한 사전분포이다. $p(\mathbf{D}|H_k)$는 모형 H_k하에서의 자료 $\mathbf{D}$의 주변분포 또는 주변우도함수이고 $p(\mathbf{D}|\theta_k, H_k)$는 θ_k의 값이 주어졌을때 $\mathbf{D}$의 분포함수이다. 식(2.4)의 형태는 일반적으로 적분이 용이하지 않은 경우가 많기 때문에 이를 계산하기 위한 방법에는 근사식을 이용한 Laplace 방법, Schwarz 기준법, Monte Carlo 적분방법, 주표본(importance sampling) 방법 등이 있다.

베이즈요인의 결과값은 비교되는 두 모형 중에서 자료에 적합한 하나의 모형을 선택할 수 있는 근거를 제시해준다. Jeffreys(1961)는 베이즈요인값에 따라 모형선택의 기준을 표 2.1과 같이 제시하였다. 이를 기준으로 두 모형 중에서 자료에 적합한 하나의 모형을 선택할 수 있다.

표 2.1: 베이즈요인을 이용한 모형선택 기준

$2\log_e(B_{10})$	B_{10}	H_0 기각에 대한 판정
$0 \sim 2$	$1 \sim 3$	모호함
$2 \sim 6$	$3 \sim 20$	확실히 기각
$6 \sim 10$	$20 \sim 150$	강하게 기각
> 10	> 150	매우강하게 기각

2.2. Metropolis-Hastings 알고리즘

베이지안 추론에서 각 추정값들은 모수의 사후분포에 대한 기대값의 형식으로 표현되어 질 수 있다. 따라서 X를 확률변수라고 하고, $\pi(x)$를 사후분포라고 하면 일반적인 베이지안 추론의 형태는 식(2.5)와 같이 표현할 수 있다.

$$E(f(X)) = \frac{\int f(x)\pi(x)dx}{\int \pi(x)dx} \quad . \tag{2.5}$$

여기서 고차원의 적분이 요구되는 경우가 많으므로 $E(f(X))$를 계산하기 위해서 다음과 같은 몬테카를로 적분 방법을 사용한다.

$$E(f(X)) \approx \frac{1}{n}\sum_{t=1}^{n} f(X_t) \quad .$$

이때 $\{X_t,\ t=1,2,\cdots,n\}$는 사후분포 $\pi(x)$에서 추출한 표본이다.

사후분포 $\pi(x)$는 일반적으로 비표준(non-standard) 형태의 함수가 많으므로 $\pi(x)$로부터 직접 $\{X_t\}$표본을 추출하기가 어렵다. 따라서 $\pi(\cdot)$를 정상분포(stationary distribution) 로 가지는 마코프연쇄(Markov chain)를 이용하여 $\{X_t\}$를 생성하는데 이 방법을 Markov chain Monte Carlo(MCMC)라고 한다. MCMC 방법을 이용하여 $E(f(X))$을 추정하기 위해서는 먼저 $\pi(x)$로 부터 $\{X_t\}$를 생성해야한다. 이 $\{X_t\}$를 생성하기 위해 전이행렬(transition matrix)을 찾아내는 Metropolis- Hastings 알고리즘(Metropolis et al., 1953; Hastings, 1970)을 이용한다.

Metropolis-Hastings 알고리즘은 먼저 t시점에서 임의의 분포 $q(\cdot|X_t)$로 부터 Y를 추출하고, 이를 이용하여 다음 상태를 나타내는 X_{t+1}를 결정하게 된다. 이때 Y는 확률 $\alpha(X_t,Y)$를 이용하여 선택된다. 여기서 $\alpha(X_t,Y)$는 아래와 같이 표현된 식으로부터 정의된다.

$$\alpha(X_t,Y)=min\left(1,\ \frac{\pi(Y)q(X_t|Y)}{\pi(X_t)q(Y|X_t)}\right) \quad .$$

즉, Y가 채택되면 $X_{t+1}=Y$가 되고 채택되지 않으면 $X_{t+1}=X_t$가 되어 연쇄는 이동하지 않게 된다.

3. 비선형회귀모형의 선택과 진단

3.1. 제안된 방법

이 절에서는 베이즈 요인의 계산과정에서 요구되는 식 (2.4)의 계산절차를 소개하고, 비선형 모형의 MCMC 과정을 통한 모수추정에 대한 수렴진단의 방법을 소개한다. 먼저 본 논문에서는 베이즈 요인의 계산과정에 사용되는 모수들에 대한 사전분포로써 다음과 같은 정보적사전분포 (informative prior)를 이용하고자 한다.

$$\theta_i \sim N(\hat{\theta}_i,\ (\sqrt{n}se(\hat{\theta}_i))^2) \quad . \tag{3.1}$$

여기서 $\hat{\theta}_i$은 최우추정량이다. 그리고 오차분산 σ^2에 대한 사전분포는 $\tau=1/\sigma^2$의 정도(precision)로 변환하여 사전분포가 넓은 범위에 걸쳐 값을 취할 수 있도록 감마분포(Gamma (0.001, 0.001)) 를 이용하기로 한다. 물론 모수 θ들에 대한 충분한 정보가 없을 때는 비정보적 사전분포를 고려하여야 하며, 베이즈 요인을 직접 계산하는데 많은 어려움이 있어 IBF(Intrinsic Bayes Fator) 또는 FBF (Fractional Bayes Factor) 등을 이용하는 방법을 생각할 수 있으나 본 연구에서는 식(3.1)의 정보적 사전분포를 고려하기로 한다.

위의 사전분포를 이용하여 베이즈요인값을 계산하기 위해서는 식(2.4)에서 언급한 다음의 적분에 대한 계산이 요구된다.

$$I=\int p(\mathbf{D}|\theta,H)\pi(\theta|H)d\theta \quad . \tag{3.2}$$

식(3.2)의 계산은 일반적으로 고차원의 적분으로 표현되거나 복잡한 형태를 따르고 있어 적분이 용이하지 않은 경우가 많다. 따라서 본 논문에서는 이 문제를 해결하기 위한 방법으

로 식(3.2)에 대한 Laplace 근사(Tierney와 Kadane, 1986)를 실시한 후, 모수에 대한 추정치로 MCMC과정을 통해 생성되는 난수에 기초한 Laplace-Metropolis 추정(Raftery, 1995)과정을 제시한다. 이 방법에 의하여 식(3.2)의 I는 다음과 같이 근사될 수 있다.

$$\hat{I} = (2\pi)^{d/2}|\tilde{\Sigma}|^{1/2}p(\mathbf{D}|\tilde{\theta}, H)\pi(\tilde{\theta}|H) \quad . \tag{3.3}$$

여기서 d는 θ의 차원을 나타내며, $\tilde{\theta}$는 사후분포의 최빈값(mode)이고, $\tilde{\Sigma}$은 모수 θ들에 대한 분산-공분산행렬로써 $\tilde{\Sigma} = (-\mathbf{D}^2\tilde{l}(\tilde{\theta}))^{-1}$이 된다. 이때, $\mathbf{D}^2\tilde{l}(\tilde{\theta})$은 2차 도함수인 헤시안 행렬(Hessian Matrix)이고, $\tilde{l}(\tilde{\theta}) = \log(p(\mathbf{D}|\theta, H)\pi(\theta|H))$이다.

또한, 식(3.3)의 추정량에 사용된 $\tilde{\theta}$와 $\tilde{\Sigma}$의 계산과정은 다음과 같다. 먼저 2절에서 소개한 Metropolis-Hastings 알고리즘을 이용하여 사후분포로부터 난수들을 생성한다. 다음으로 생성된 난수들을 이용하여 계산에 요구되는 $\tilde{\theta}$와 $\tilde{\Sigma}$의 두 값을 구하면 된다. 사후분포의 최빈값인 $\tilde{\theta}$를 구하기 위한 가장 간단한 방법은 $p(\mathbf{D}|\theta^{(i)})\pi(\theta^{(i)})$의 값을 최대화하는 $\theta^{(i)}$로 추정하는 것이고 이때 각각 추출된 $\theta^{(i)}$에서 우도함수 값의 계산이 요구된다. 분산-공분산 행렬 Σ에 대한 추정은 $\tilde{\Sigma}$가 $n \to \infty$가 됨에 따라 근사적으로 사후분포의 분산-공분산 행렬과 같게 된다는 사실을 근거로 $\tilde{\Sigma}$가 사후분포의 모의실험한 결과로부터 추정된 사후분포의 분산-공분산 행렬을 이용하기로 한다.

위에서 소개된 절차에 따라 계산된 추정량 $\tilde{\theta}$와 $\tilde{\Sigma}$을 이용하여 식(2.3)의 베이즈요인을 구할 수 있다. 이를 이용하여 실례를 통한 최적의 비선형회귀모형을 선택하는 과정에 대해서는 4절에서 자세히 다루게 될 것이다. 또한 위의 추정과정에서 사용된 MCMC 과정으로부터의 추정량에 대한 수렴진단과정은 다음절에서 다루었다.

3.2. 수렴진단

이 절에서는 3.1절에서 제안된 Laplace-Metropolis 추정량의 계산과정에서 모수의 Metropolis-Hastings 알고리즘을 통한 추정과정과 이들 모수의 수렴과정에 대한 진단법에 대해 다루었다. 본 논문에서 사용한 진단 방법으로는 최근에 가장 많이 사용되는 Gelman과 Rubin(1992)의 방법과 Raftery와 Lewis(1992)의 방법을 사용하기로 한다.

Gelman과 Rubin(1992)의 방법은 2개 이상의 병렬연쇄(pararell chains)를 통해 수렴과정을 진단하는 방법으로, 각 연쇄(chain)는 실제 사후분포에 대하여 과대산포(over-dispersion)를 갖는 각기 다른 초기값을 이용하여 수행한다. 이 방법은 각 연쇄에 대해 연쇄내 분산(within chain variance)과 연쇄간 분산 (between chain variance)을 비교하는데 기초를 두고 있다. 이 방법의 측도는 다음과 같다.

$$\sqrt{\hat{R}} = \sqrt{\frac{\widehat{var}^{+}(\theta|\mathbf{D})}{W}} \quad .$$

여기서 $\widehat{var}^{+}(\theta|\mathbf{D}) = \frac{n-1}{n}W + \frac{1}{n}B$이고 W는 연쇄내 분산, B는 연쇄간 분산을 나타낸다. 만약 연쇄가 무한대로 진행됨에 따라 이 측도의 값이 1에 가까워지면 두 연쇄에 의해 생성된 θ들이 동일한 분포로 수렴한다는 것을 나타낸다.

Raftery와 Lewis(1992) 방법은 한 개의 연쇄에 대해서 적용하는 방법으로 다음과 같은 측도 I(dependence factor)를 이용하여 수렴진단을 실시한다.

$$I = \frac{M+N}{N_{min}} \quad .$$

여기서 M은 제거(burn-in)해야하는 반복수이고, N은 전체반복수이며, N_{min}은 θ들이 독립인 경우에 분위수를 계산하기 위해서 필요한 최소 반복수이다. 그러므로 I는 연쇄안에 있는 표본 θ들이 엄밀하게 독립은 아니기 때문에 독립표본으로 간주되기 위해 요구되는 반복수의 증가여부를 나타내는 측도이다. $I > 1$이면 표본들이 서로 상관관계가 높다는 것을 나타내고, $I > 5$이면 모형의 재모수화(reparameterization)를 통하여 연쇄내의 상관관계를 줄임으로써 독립표본으로의 수렴속도에 대한 개선이 요구됨을 나타낸다.

4. 실제자료분석 및 모의실험

4.1. 목초의 재성장량 자료분석

4.1.1. 모형선택

이 절에서는 Ratkowsky(1983)가 목초의 성장량을 측정하고 난 이후, 시간(time)의 변화에 따라 다시 목초의 재성장량(yield)을 관측한 표4.1의 자료를 가지고 베이즈요인을 이용하여 최적모형을 선택하고 이 모형에 대한 수렴진단을 실시하였다.

표 4.1: 시간의 변화에 따른 목초의 재성장량자료 (단위:알려져있지않음.)

시간(time)	9	14	21	28	42	57	63	70	79
성장량(yield)	8.93	10.8	18.59	22.33	39.35	56.11	61.73	64.62	67.08

이 자료에서 시간 x_i에 따라 재성장량 Y_i는

$$Y_i = f(x_i, \theta) + \epsilon_i, \quad \epsilon_i \sim N(0,\ \sigma^2)$$

와 같이 표현된다. 여기서 Ratkowsky(1989)가 제안한 비선형회귀모형의 다음 4가지 함수 $f(x,\theta)$ 를 이용하여 베이지안 방법으로 최적모형을 선택한다.

$$[1] f(x,\theta) = \theta_1 - \theta_2 \exp(-\exp(\theta_3 + \theta_4 \log x))$$
$$[2] f(x,\theta) = \theta_1 \exp(\theta_2/(x+\theta_3))$$
$$[3] f(x,\theta) = \theta_1 \exp(-\exp(\theta_2 - \theta_3 x))$$
$$[4] f(x,\theta) = \theta_1 + \theta_2/(1+\exp(\theta_3 - \theta_4 x))$$

본 연구에서는 표본을 100,000개 추출하고 그 표본들 중에서 초기 생성자료 20,000개를 제거(burn-in)하여 사후분포로부터 전체 80,000개의 θ들을 추출하였다. 이 표본들을 가지

고 식(3.3) 의 Laplace-Metropolis 추정방법을 이용하여 베이즈요인 값을 계산하고 이것을 통해 최적모형을 선택한 결과를 표4.2에 수록하였다. 표4.2의 결과를 표2.1의 모형선택 기준에 따라 살펴보면 목초의 재성장량 자료에 적합시킨 4가지 모형들 중에서 [모형1]과 [모형2]가 선택되었다. 최종적으로 모수 축약의 원칙을 적용하면 모수의 개수가 보다 작은 [모형2]를 최적모형으로 결정할 수 있다.

표 4.2: 베이즈요인의 추정과 모형선택

모형	베이즈요인	베이즈요인의추정값	모형선택
[모형1]과 [모형2]	B_{21}	2.2	[모형1]과 [모형2]
[모형1]과 [모형3]	B_{13}	129831.8	[모형1]
[모형1]과 [모형4]	B_{14}	88139.3	[모형1]
[모형2]과 [모형3]	B_{23}	289309.3	[모형2]
[모형2]과 [모형4]	B_{24}	196404.2	[모형2]
[모형3]과 [모형4]	B_{34}	1.5	[모형3]과 [모형4]

즉,

$$[2]\ f(x,\theta) = \theta_1 \exp(\theta_2/(x+\theta_3))$$

가 최적모형으로 선택된 것이다.

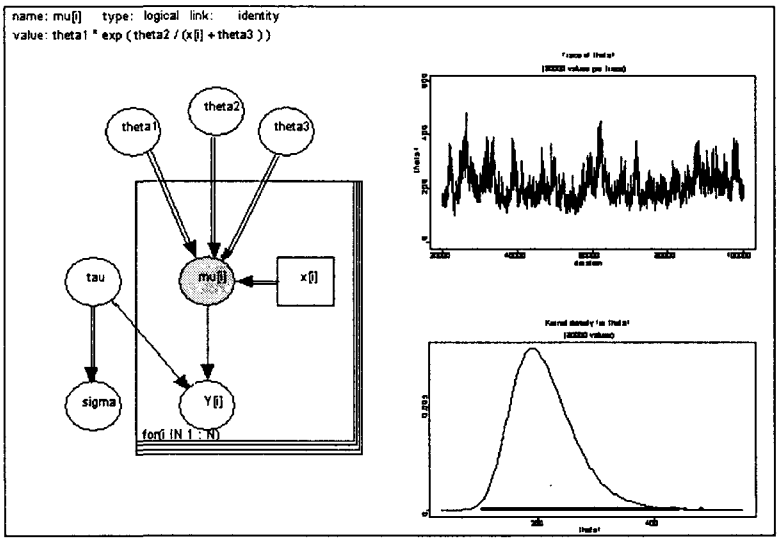

그림 4.1: MCMC를 이용한 모수의 추정과정과 결과

참고로 위의 그림4.1은 MCMC를 이용한 θ_1의 추정과정과 그 결과를 나타낸 것이다. 그림의 좌측은 모수의 추정과정을 시각적으로 모형화한 그래픽모형으로 기호 →은 확률적 연

결을 나타내고 기호 ⇒은 논리적 연결을 의미한다. 또한 그림의 우측 첫 번째 결과는 MCMC 과정을 통해 사후분포로부터 80000개 발생한 θ_1 난수의 생성 결과를 나타내고, 두 번째 그림은 발생된 난수로부터 θ_1의 분포에 대한 커널 밀도함수 추정을 한 결과이다. 본 논문에서는 지면관계상 θ_1의 추정 결과만 수록하였다.

4.1.2.수렴진단

먼저 Gelman과 Rubin(1992)의 방법을 이용하여 최적모형으로 선택된 [모형2]의 모수들에 대하여 수렴진단을 실시한 결과는 그림4.2와 같다. 이 방법은 Splus에서 제공하는 coda() 함수를 이용하여 실행한 결과로써 두 연쇄에 의해 생성된 θ들의 두 분위수 (50%와 97.5%) 값들에 대한 Shrink Factor의 값이 출력된다. 각각의 모수의 두 분위수에 대한 Shrink Factor의 값이 모두 1의 값으로 근사할 때, 사후분포로부터 MCMC 과정을 통한 모수들의 수렴이 제대로 이루어지고 있음을 알 수 있다. 즉, 그림4.2의 결과로부터 각 모수에 대한 사후분포로부터의 난수생성이 효과적임을 알 수 있다.

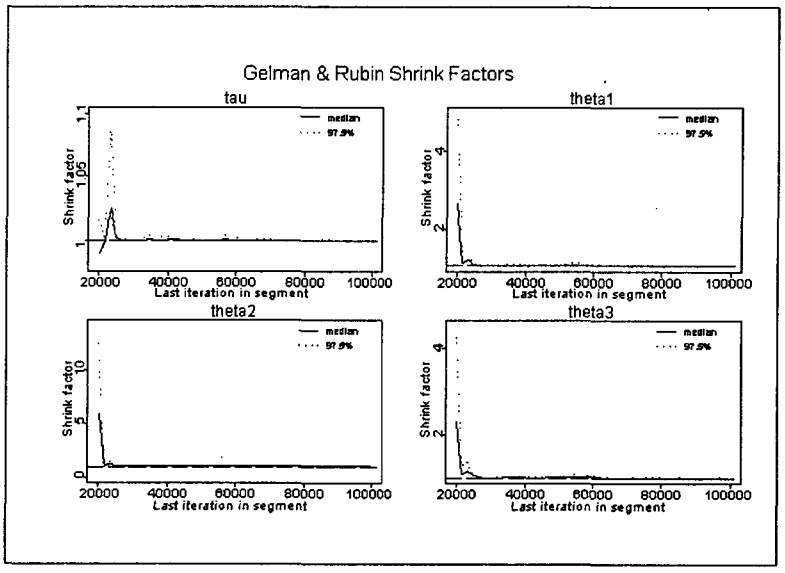

그림 4.2: [모형2]에 대한 Gelman과 Rubin의 수렴진단

또한 Gelman 과 Rubin(1992)의 수렴진단 결과를 수치적으로 나타내면 표4.3과 같이 나타나며, 그림4.2의 진단결과와 동일함을 알 수 있다.

다음은 Raftery와 Lewis(1992)의 수렴진단 결과를 소개하면 표4.4와 같다. 이 경우 생성된 모수값들에 대한 I(dependence factor)값이 모두 5이상으로 나타나고 있으므로 모형의 재모수화를 통하여 연쇄내의 상관관계를 줄임으로써 독립표본으로의 수렴속도를 개선해야할 필요성이 있음을 보이고 있다.

표 4.3: [모형2]의 Gelman과 Rubin의 진단결과

모수	[모형2]에 대한 Shrink Factor	
	50%	97.5%
τ	1	1
θ_1	1	1
θ_2	1	1
θ_3	1	1

표 4.4: [모형2]의 Raftery와 Lewis의 진단결과

모수	모형2			
	M	N	N_{min}	I
τ	15	21415	3746	5.72
θ_1	260	255640	3746	68.2
θ_2	900	974988	3746	260
θ_3	341	348471	3746	93

4.2. 모의실험

이 절에서는 참모형으로부터 인위적인 자료를 생성하고 본 연구에서 제시한 방법이 효과적으로 참모형을 선택하는지에 대한 추가적인 모의실험을 수행하였고, 아울러 고전적인 모형선택법인 불일치측도(discrepancy measure)를 이용한 결과와도 비교분석을 실시하였다.

먼저 다음의 식(4.1)의 모형으로부터 표4.5의 모의실험자료를 생성하였다.

$$y_{ij} = 1 + e^{2x_i}(1 + 3x_i) + e_{ij}, \qquad i = 1, 2, \cdots, 10 \quad j = 1, 2, 3. \tag{4.1}$$

여기서 $x_i = (i-6)/5$이고, $e_{ij} \sim N(0, 2)$이다. 본 모의실험에서는 다음의 4가지 비선형 회귀모형을 대상으로 최적모형 선택 과정을 수행하였다.

[1] $f(x, \theta) = \theta_1 e^{\theta_2 x} + \theta_3$

[2] $f(x, \theta) = \theta_1 x^2 + \theta_2 x + \theta_3$

[3] $f(x, \theta) = e^{\theta_1 x}(\theta_2 + \theta_3 x) + \theta_4$

[4] $f(x, \theta) = \theta_1 x^3 + \theta_2 x^2 + \theta_3 x + \theta_4$

모의실험을 수행한 결과 표4.6에서와 같이 [모형3]이 최적모형으로 선택되어 참모형인

식(4.1) 과 일치하였다. 따라서 본 연구에서 제시한 방법이 효과적으로 참모형을 선택한다는 것을 알 수 있다.

표 4.5: 모의실험자료

i	y_{i1}	y_{i2}	y_{i3}
1	2.21	0.18	0.98
2	1.74	0.52	0.57
3	1.61	2.47	-1.46
4	-2.77	1.89	2.66
5	-0.70	-0.40	0.66
6	1.64	1.95	2.71
7	4.41	3.46	4.54
8	8.34	5.42	5.55
9	10.86	10.21	11.54
10	19.32	19.15	17.83

표 4.6: 베이즈요인의 추정과 모형선택

모형	베이즈요인	베이즈요인의추정값	모형선택
[모형1] 과 [모형2]	B_{12}	10114.0	[모형1]
[모형3] 과 [모형1]	B_{31}	761.9	[모형3]
[모형1] 과 [모형4]	B_{14}	32133.0	[모형1]
[모형2] 과 [모형3]	B_{23}	77055.0	[모형3]
[모형2] 과 [모형4]	B_{24}	31.5	[모형4]
[모형3] 과 [모형4]	B_{34}	24481.0	[모형3]

또한 고전적 모형선택법으로 사용되고 있는 식(4.2)의 가우스 불일치측도 기준을 이용하여 최적모형을 선택하였다.

$$\sum_{ij}(\bar{y}_{i.} - h(x_i, \hat{\theta}))^2 + 2p\hat{\sigma}^2 \quad . \tag{4.2}$$

여기서 $h(x_i, \hat{\theta})$은 근사적 평균함수이고 p는 모수의 개수가 되며 $\hat{\sigma}^2$은 평균제곱오차(MSE)로 다음과 같이 계산된다.

$$\hat{\sigma}^2 = \sum_{ij}(y_{ij} - \bar{y}_{i.})^2/(J_. - I) \quad .$$

이때 J은 전체 자료의 개수이고 I는 수준의 개수를 나타낸다.

가우스 불일치측도를 이용한 최적모형의 선택결과는 표4.7에서와 같이 불일치측도 기준값이 가장 작은 [모형1]이 최적모형으로 선택되었다. 그러나 [모형3]도 [모형1]의 불일치측도 기준값과 거의 차이가 나지 않으므로 적합한 모형이라 말할 수 있다. 이는 베이즈 요인에 기초한 베이지안 모형선택 결과와 유사하며 다만 베이지안 방법의 경우 사전정보를 이용한다는 측면에서 의미가 있다고 말할 수 있다.

표 4.7: 가우스 불일치측도 기준값

모형	모수의수 (p)	$\sum_{ij}(\bar{y}_{i\cdot} - h(x_i, \hat{\theta}))^2 + 2p\hat{\sigma}^2$
[모형1]	3	30.041
[모형2]	3	65.808
[모형3]	4	30.973
[모형4]	4	32.492

5. 결론

본 논문에서는 베이지안 기법에 기초한 최적의 비선형회귀모형의 선택법을 다루었다. 베이즈요인에 기초한 이 방법은 고전적인 모형선택법에 비해 사전정보를 이용하거나 소표본 및 비내포 모형에도 효과적으로 사용될 수 있다는 측면에서 대단히 효율적인 방법이라 할 수 있다. Laplace 근사와 Metropolis-Hastings 알고리즘을 이용한 MCMC 시뮬레이션 방법에 기초한 이 방법은 고차원이거나 복잡한 유형의 모형에서 모수 추론의 어려움을 해결해 주며 고전적인 모형선택과정에서 요구되는 복잡한 근사식의 유도가 불필요하다는 장점을 가진다.

본 논문에서는 베이즈요인의 계산과정에서 모수에 대한 사전분포로써 정보적 사전분포(informative prior)를 이용하였다. 모수에 대한 사전정보가 충분하지 않은 경우에는 비정보적 사전분포 (noninformative prior)를 사용하는 방법에 대해 추가적인 연구가 필요하다. 이 경우에는 IBF(Intrinsic Bayes Factor) 또는 FBF(Fractional Bayes Factor)를 이용하여 베이즈요인에 대한 효과적인 추정을 실시하는 것이 바람직하다.

참고문헌

[1] Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences(with discussion). *Statistical Science*, 7, 457-511.

[2] Hastings, W. K. (1970). Monte Carlo Sampling Methods using Markov Chains and their applications. *Biometrika*, 57, 97-109.

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

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

[5] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21, 1087-1091.

[6] Raftery, A. E. and Lewis, S. (1992). How Many Iterations in the Gibbs Sampler?. *In Bayesian Statistics 4*(eds. J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M.Smith), Oxford: Oxford University Press, 763-773.

[7] Raftery, A. E. (1995). Hypothesis Testing and Model Selection via Posterior Simulation. *Markov Chain Monte Carlo In Practice* (by W.R. Gillks, D.J. Spiegelhalter and S. Richardson, eds.), London: Chapman & Hall.

[8] Ratkowsky, D. A. (1989). *Handbook of Nonlinear Regression Models*, M. Dekker, New York.

[9] Ratkowsky, D. A. (1983). *Nonlinear Regression Modeling*, M. Dekker, New York.

[10] Tierney, L. and Kadane, J. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, 81, 82-86.

[2000년 9월 접수, 2002년 3월 채택]

Bayesian Model Selection and Diagnostics for Nonlinear Regression Model *

Jong-Hwa Na[1] Jeong-Suk Kim [2]

ABSTRACT

This study is concerned with model selection and diagnostics for nonlinear regression model through Bayes factor. In this paper, we use informative prior and simulate observations from the posterior distribution via Markov chain Monte Carlo. We propose the Laplace approximation method and apply the Laplace-Metropolis estimator to solve the computational difficulty of Bayes factor.

Keywords: MCMC; Metropolis-Hastings algorithm; Laplace-Metropolis;estimator;Bayes factor; Convergence diagnostics.

* This work was supported by grant No.(2001-1-10400-017-1) from the Basic Research Program of the Korea Science & Engineering Foundation.

1) Associate Professor, Department of Statistics, Chungbuk National University.
E-mail : cherin@cbucc.chungbuk.ac.kr

2) Graduate Student, Department of Statistics, Chungbuk National University.
E-mail :. chastity@trut.chungbuk.ac.kr