응용통계연구 (2010)
23(6), 1169–1178
DOI: 10.5351/KJAS.2010.23.6.1169

불균형 자료에서 AIC를 이용한 선형혼합모형 선택법의 효율에 대한 모의실험 연구

이용희[1]

[1]서울시립대학교 통계학과

(2010년 9월 접수, 2010년 10월 채택)

요 약

본 논문은 불균형 자료에서 선형혼합모형에 적용되는 Akaike Information Criterion(AIC)의 효율에 대한 연구이다. Vaida와 Balanchard (2005)에 의해 제안된 cAIC(conditional AIC)는 mAIC(marginal AIC)가 임의효과의 예측에 대한 불확실성을 모형선택에서 반영하지 못하는 단점을 극복할 수 있는 방법이다. cAIC에 대한 이론적인 성질과 확장은 Liang 등 (2008)과 Greven과 Kneib (2010)에 의하여 연구되었다. cAIC의 형태는 자료의 구조에 영향을 받지는 않지만 선형혼합모형에서 모수의 추정 효율은 자료의 불균형의 정도에 따라 많은 영향을 받는 것이 알려져 있다. 기존의 연구에서 실시한 모든 모의실험이 자료가 균형인 경우에만 실행되어 자료의 불균형이 AIC에 근거한 혼합모형 선택 방법의 효율에 어떤 영향을 미치는지 알려져 있지 않다. 본 논문은 자료의 불균형이 모형선택 방법의 효율에 미치는 영향을 모의실험을 통하여 알아보았다. 자료의 불균형이 심해짐에 따라 AIC에 근거한 모형선택방법은 복잡한 모형을 선택하는 경향이 낮아짐을 보였다.

주요용어: 선형혼합모형, 자료의 불균형, AIC, 모형선택.

1. 서론

선형혼합모형(Linear Mixed-Effects Models)은 서로 독립이 아닌 다양한 형태의 상관관계를 가지는 자료들에 적용할 수 있는 매우 유용한 통계적 모형이다. 예로 반복측정 자료나 군집 자료에서는 관측값들이 서로 독립이 아니며 또한 동일한 분포를 가지지 않기 때문에 단순한 선형모형을 적용할 수 없다. 이러한 경우 선형혼합모형은 임의효과(random effects)를 이용하여 자료들 간에 다양한 상관관계를 설명할 수 있다. 선형혼합모형하에서의 통계적 추론은 분산분석(ANOVA; Analysis of Variance)의 제곱합을 이용한 적률추정방법에서 시작되었으며 Hartley와 Rao (1967)에 의해 우도함수방법이 제안되면서 다양한 추정 방법과 그 계산법이 개발되어 왔다 (Harville, 1977; Searle 등, 1992; Jiang, 2009). 특별히 자료가 불균형(unbalanced data)인 경우에는 대부분의 추정량이 자료가 균형일 때 나타나는 최적성질을 만족하지 않는 것이 알려져 있으며 자료가 불균형일 때 생기는 추정량 효율의 저하를 극복하기 위하여 많은 연구가 수행되어 왔다 (Khuri 등, 1998).

주어진 자료에 대한 통계적 모형들을 고려할 때 가장 적절한 모형을 선택하기 위한 방법에는 변수선택법(variable selection), 통계적 검정(hypothesis testing), 정보기준(Information Criteria) 등을 이용한 다양한 방법들이 있으며 이들 중 가장 많이 쓰이는 기준으로 Akaike Information Criterion(AIC)이 있

이 연구는 2009학년도 서울시립대학교 교내학술연구비 지원에 의한 연구되었음.
[1](130-743) 서울시 동대문구 전농동 90 시립대길 13, 서울시립대학교 통계학과, 교수. E-mail: ylee@uos.ac.kr

다 (Akaike, 1973). AIC는 Kullback-Leibler divergence에 의해 정의되며 통계 모형이 자료를 얼마나 잘 설명하는지에 대한 정도를 나타내는 우도함수의 값과 모형의 복잡한 정도를 나타내는 값의 결합으로 구성된다. 일반적으로 정규분포 가정을 하고 서로 독립인 관측값에 대한 선형모형 $y = X\beta + e$에서 AIC는 다음과 같이 정의된다.

$$\text{AIC} = -2l + 2(p+1), \tag{1.1}$$

여기서 l은 정규분포 $N(X\beta, \sigma^2 I)$에 대한 로그우도함수이며 p는 고정모수 벡터 β의 차원이다. 모형의 복잡한 정도를 나타내는 AIC 보정항은 $2(p+1)$으로 주어지며 이는 모수벡터의 차원을 반영한다. 즉 β의 차원에 오차항의 분산 σ^2 한 개를 더하면 $p+1$이 된다. 보정항에서 알 수 있듯이 β의 차원이 늘어나면 모형의 복잡성이 증가하여 모형의 선택의 기준에 대한 벌칙으로 반영된다. 선형모형에 대한 AIC는 선형혼합모형으로 확장될 수 있으며 모형에서 관심 있는 양에 따라서 두 가지 서로 다른 형태로 확장된다. 만약 선형혼합모형에서 주요 관심사가 관측값의 평균과 분산에 대한 모수의 추정이라면 임의효과를 적분한 뒤 얻어지는 주변분포함수에 근거한 mAIC(marginal AIC)를 이용할 수 있으며 이는 식 (1.1)에 주어진 AIC의 자연스러운 확장이다. 반면 선형혼합모형에서 임의효과 예측에 관심이 있다면 임의효과가 주어진 조건부확률분포에 근거한 cAIC(conditional AIC)를 이용할 수 있다 (Vaida와 Blanchard, 2005; Liang 등, 2008).

선형혼합모형에서 mAIC와 cAIC를 이용한 모형의 선택에 대한 모든 연구들은 모의실험이 자료의 구조가 균형인 경우에만 실시되었다. AIC의 형태는 균형인 경우와 자료가 불균형인 경우에 동일하지만 불균형 자료인 경우 임의효과의 분산에 대한 추정량의 효율이 균형 자료인 경우와 다르다고 알려져 있다. 따라서 AIC에 의거한 선형혼합모형의 선택법이 균형 자료와 불균형 자료에서 어떤 차이를 보이는지에 대한 연구가 필요하다. 본 논문에서는 이러한 차이를 모의실험을 통하여 알아보고자 한다. 2장에서 선형혼합 모형과 AIC에 근거한 모형의 선택법을 살펴보고 3장에서 불균형 자료를 이용한 모의실험으로 모형선택 방법의 효율이 자료의 불균형에 따라 얼마나 영향을 받는지 알아본다.

2. 선형혼합모형과 AIC에 의한 모형의 선택

2.1. 선형혼합모형

선형혼합모형은 임의효과를 이용하여 자료들 간에 상관관계와 분포의 이질성을 설명할 수 있으며 일반적인 형태는 다음과 같이 나타낼 수 있다.

$$y = X\beta + Z\alpha + e, \tag{2.1}$$

여기서 y는 $n \times 1$ 관측값 벡터이고 X는 $n \times p$ 행렬로 고정효과 $\beta' = (\beta_1, \ldots, \beta_p)$에 대한 설계행렬이다. Z는 $n \times r$ 행렬로 임의효과 $\alpha' = (\alpha_1, \ldots, \alpha_r)$에 대한 설계행렬이다. 임의효과 α는 정규분포 $N(0, D(\theta))$를 따르는 확률 변수이며 θ는 임의효과의 분산 $D = D(\theta)$에 대한 모수이다. 오차 벡터 e는 서로 독립이며 정규분포 $N(0, \sigma^2 I)$를 따르는 오차들로 이루어져 있다. 이러한 모형 하에서 관측치의 분산은 다음과 같이 나타내어진다.

$$\text{Var}(y) \equiv V = ZDZ' + \sigma^2 I. \tag{2.2}$$

따라서 선형혼합모형은 자료의 구조에 적합한 임의효과를 고려하여 관측치 간의 다양한 상관관계를 설명할 수 있는 매우 유용한 모형이다.

2.2. 선형혼합모형에서의 AIC

AIC(Akaike Information Criterion)는 Akaike (1973)에 의해 제안된 모형선택의 기준으로서 자료를 생성하는 실제 모형과 자료를 분석하기 위해 고려된 모형의 거리를 나타내는 Kullback-Leibler divergence(K-L divergence)를 반영하는 모형선택의 기준이다. 자료를 생성하는 실제모형을 $f_t(y)$라 하고 분석에서 사용된 모형의 집합을 $\{f_\psi(y)|\psi \in \Psi\}$라 하면 K-L divergence는 다음과 같이 정의된다.

$$\mathrm{KL}(f_t, f_\theta) = E_t(\log f_t) - E_t(\log f_\psi).$$

K-L divergence에서 $E_t(\log f_t)$는 고려된 모형과 상관이 없는 상수이기 때문에 $-E_t(\log f_\psi)$의 값이 모형의 선택에서 추정하고자 하는 양이며 그 값이 작을수록 실제 모형과 가깝다. 따라서 주어진 모형이 실제 모형에 가까운 정도를 나타내는 양으로서 Akaike Information을 다음과 같이 정의한다.

$$\mathrm{AI} = -2E_t(\log f_\psi).$$

모수 ψ는 보통 최대우도 추정량 $\hat{\psi} = \hat{\psi}(y)$로 추정되며 Akaike Information에 대한 추정값이 선형모형 $y = X\beta + e$에서는 식 (1.1)의 AIC이며 이는 우도함수와 모형의 복잡성이 결합한 형태이다. 선형혼합모형에 대한 AIC는 모형에서 관심 있는 양에 따라서 두 가지 서로 다른 형태로 확장된다.

선형혼합모형에서 평균과 분산성분에 대한 모수의 추정에 관심이 있는 경우 우도함수를 임의효과에 대하여 적분한 주변분포함수를 이용한다. 임의효과가 주어진 조건부분포를 $f(y\,|\,\beta,\, \sigma^2, \alpha)$이라 하고 임의효과의 분포를 $g(\alpha|\theta)$라 하면 주변로그우도함수 $l = l(\beta, \theta, \sigma^2)$는 다음과 같이 주어진다.

$$l(\beta, \theta, \sigma^2) = \log \int f\left(y|\beta, \sigma^2, \alpha\right) g(\alpha|\theta) d\alpha.$$

관측벡터 y가 선형혼합모형 (2.1)을 따른다고 가정하면 임의효과에 대하여 적분한 주변분포함수는 정규분포 $N(X\beta,\, ZDZ' + \sigma^2 I)$로 주어지며 이 경우에 AIC는 mAIC로 확장될 수 있으며 다음과 같이 정의할 수 있다.

$$\mathrm{mAIC} = -2l + 2K = -2l\left(\hat{\beta}, \hat{\theta}, \hat{\sigma}^2\right) + 2K. \tag{2.3}$$

mAIC의 로그우도함수 l에서 β는 선형모형에서의 평균에 대한 모수이고 θ는 임의효과 α의 분산에 대한 모수, σ^2는 오차항의 분산에 대한 모수이다. 여기서 모형의 복잡성을 나타내는 측도인 K는 다음과 같이 주어지며 주어진 모형에 대한 모수(parameters)의 수와 같다.

$$K = \dim(\beta) + \dim(\theta) + 1.$$

선형혼합모형에서는 모수의 추정뿐만 아니라 임의효과의 예측도 중요한 문제이다. 따라서 이러한 경우에 mAIC처럼 임의효과에 대한 예측을 고려하지 않는 양은 모형선택의 기준으로서 제한적이다. 임의효과에 대한 예측을 고려하는 모형선택의 기준이 cAIC이며 다음과 같이 정의된다.

$$\mathrm{cAIC} = -2l + 2K = -2\log f\left(y\,|\,\hat{\beta}, \hat{\theta}, \hat{\sigma}^2, \hat{\alpha}\right) + 2K. \tag{2.4}$$

cAIC에서 로그우도함수는 임의효과의 예측값 $\hat{\alpha}$가 주어진 조건부확률분포로부터 얻어진다. cAIC에서의 모형의 복잡도 K를 추정하는 방법은 선형혼합모형에서 분산성분 모수 θ과 오차항의 분산 σ^2을 알고 있다는 가정 하에서 Vaida와 Blanchard (2005)에 의하여 제안되었으며 $K = \rho$로 하면 ρ는 다음과 같이 정의된다.

$$K = \rho = \mathrm{tr}\left\{ \begin{pmatrix} X & Z \end{pmatrix} \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + D_*^{-1} \end{pmatrix}^{-1} \begin{pmatrix} X' \\ Z' \end{pmatrix} \right\} \equiv \mathrm{tr}(H_1), \tag{2.5}$$

여기서 $D_* = D(\theta)/\sigma^2$이며 모형의 복잡도를 나타내는 ρ는 관측값의 조건부 예측값 $\hat{y} = \hat{E}(y|\hat{\beta}, \hat{\theta}, \hat{\alpha})$를 관측값 벡터의 선형함수 $\hat{y} = H_1 y$로 나타낼 때 사영행렬(projection matrix)과 유사한 H_1의 trace이다. 따라서 식 (2.5)에 나타난 ρ는 선형모형의 복잡성에 대한 측도 $\rho = \mathrm{tr}[X(X'X)^{-1}X'] = p$를 선형혼합모형으로 확장한 것이다. Vaida와 Blanchard (2005)는 더 나아가 오차항의 분산 σ^2은 모르고 분산성분의 모수 θ를 안다고 가정했을 때 Akaike Information에 대한 불편추정량 cAIC를 위한 K를 다음과 같이 구하였다.

$$K = \frac{n(n-p-1)}{(n-p)(n-p-2)}(\rho+1) + \frac{n(p+1)}{(n-p)(n-p-2)}. \tag{2.6}$$

더 나아가 분산성분의 모수 θ를 추정해야 하는 경우에 대한 K 값은 Liang 등 (2008)과 Greven과 Kneib (2010)에 의하여 제안되었다. Liang 등 (2008)에서는 분산성분의 모수 θ의 추정에 대한 불확실성까지 포함하는 수정항 K를 다음과 같이 제안하였다.

$$K = \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i},$$

여기서 $\hat{y}_i$는 조건부 예측값 벡터 $\hat{y} = X\hat{\beta} + Z\hat{\alpha}$의 i번째 원소이다. Liang 등 (2008)의 수정항 K를 계산할 때 수치적 미분법을 사용할 수 있으나 자료의 개수만큼의 미분이 필요하므로 계산 시간이 많이 걸리는 단점이 있다. Greven과 Kneib (2010)은 Liang 등 (2008)의 수정항 K를 음함수 미분법을 통하여 구체적인 공식을 제시하였으나 그 형태가 매우 복잡하다. 또한 Greven과 Kneib (2010)은 주변분포가 모형선택의 목표로 정해진 경우 대해서도 mAIC가 Akaike Information에 대한 불편추정량이 아니라는 사실을 보였다.

3. 불균형 자료에 대한 모의실험

자료가 불균형인 경우에 분산성분에 대한 추정의 효율은 균형 자료인 경우와 다르다고 알려져 있다 (Searl 등, 1992; Khuri 등, 1998). AIC의 형태는 균형인 경우와 동일하지만 AIC도 분산성분 추정량의 함수이므로 그 효율은 자료의 구조에 영향을 받을 것으로 예상되어진다. 따라서 선형혼합모형의 선택법이 불균형 자료일 경우 균형 자료에 비교하여 어떤 다른 점이 있는지에 대한 연구가 필요하다. 균형 자료와 불균형 자료에서의 모수 추정의 효율 차이는 이론적으로 비교하는 방법이 가능하지만 (Khuri 등, 1998), 모형선택에서 차이는 모의실험을 통하여 비교하는 것이 쉽고 간단한 방법이다.

자료의 불균형이 모형의 선택법에 미치는 영향을 알아보기 위하여 다양한 형태의 불균형 자료에서 상수평균모형(M_1)과 임의절편모형(M_2)의 선택에 대하여 모의실험을 실시하였다. 고려된 두 모형 M_1과 M_2의 방정식은 다음과 같다.

$$M_1 : y_{ij} = \mu + e_{ij}, \qquad M_2 : y_{ij} = \mu + \alpha_i + e_{ij},$$

여기서 $i = 1, 2, \ldots, I$이고 $j = 1, 2 \ldots, J_i$이며 전체 자료 개수는 $n = \sum_i J_i$이다. 임의효과 α_i는 서로 독립이며 평균이 0이고 분산이 τ^2를 따르는 확률변수이며 오차항 e_{ij} 또한 서로 독립이고 평균이 0이고 분산이 σ^2를 따르며 임의효과와는 독립이다. 모형 M_2에서 분산성분 τ^2가 0이면 모형 M_1과 동일하다.

모형 M_1인 경우에는 mAIC와 cAIC가 동일하며 AIC는 다음과 같다.

$$\mathrm{AIC} = n\log 2\pi + n\log\hat{\sigma}^2 + \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - \hat{\mu})^2 + 2(1+1).$$

모형 M_2인 경우에는 mAIC는 관측값의 주변분포로 계산되며 만약 i번째 군집의 관측벡터를 y_i라 하면 그 주변분포는 다음과 같이 주어진다.

$$y_i \sim N(\mu 1_{J_i}, V_i), \quad V_i = \sigma^2 I_{J_i} + \tau^2 1_{j_i} 1_{j_i}', \tag{3.1}$$

여기서 1_{J_i}는 모든 원소가 1인 $J_i \times 1$ 벡터이며 I_{J_i}는 차원이 J_i인 단위행렬이다. 따라서 모형 M_2에 대한 mAIC는 다음과 같이 주어진다.

$$\text{mAIC} = n \log 2\pi + \sum_i \log \left| \widehat{V}_i \right| + \sum_i (y_i - \hat{\mu} 1_{J_i})' \widehat{V}_i^{-1} (y_i - \hat{\mu} 1_{J_i}) + 2(1+1+1). \tag{3.2}$$

식 (3.2)의 mAIC에서 모형의 복잡도에 대한 상수 $K = 3$이며 이는 평균 μ와 두 개의 분산성분 σ^2과 τ^2로 이루어진 모수의 개수로 주어지며 자료의 형태나 분산성분의 추정치와는 관계없이 변하지 않는다.

모형 M_2에 대한 cAIC는 다음과 같이 주어진다

$$\text{cAIC} = n \log 2\pi + n \log \hat{\sigma}^2 + \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \hat{\mu} - \hat{\alpha}_i)^2 + 2K, \tag{3.3}$$

여기서 임의효과의 예측치 $\hat{\alpha}_i$는 다음과 같이 주어진다.

$$\hat{\alpha}_i = \frac{\hat{\tau}^2}{\frac{\hat{\tau}^2 + \hat{\sigma}^2}{J_i}} (\bar{y}_{i\cdot} - \hat{\mu}).$$

Vaida와 Blanchard (2005)에서 제안된 cAIC (3.3)의 K는 식 (2.6)로 주어지며 이때 $p = 1$이고 ρ는 다음과 같이 추정한다.

$$\hat{\rho} = \hat{\tau}^2 \left(\frac{G_0 - G_1}{G_0} \right),$$

여기서

$$\gamma_i^2 = \frac{J_i}{J_i \hat{\tau}^2 + \hat{\sigma}^2}, \quad G_0 = \sum_i \gamma_i, \quad G_1 = \sum_i \gamma_i^2.$$

분산성분 모수 θ에 대한 추정의 불확실성을 고려하는 Greven과 Kneib (2010)의 cAIC에서 수정항 K는 임의절편모형 M_2에서 다음과 같이 주어진다.

$$K = \frac{n(n-p-1)}{(n-p)(n-p-2)} (\hat{\rho} + 1) + \frac{n(p+1)}{(n-p)(n-p-2)} + \hat{B} \hat{G} \hat{A} \hat{W} y, \tag{3.4}$$

여기서 $y' = (y_1', y_2', \ldots, y_I')$, $X = 1_n$, Z는 임의효과 α에 대한 계획행렬, $V = (I_n - \tau^2 ZZ'/\sigma^2)$, $A = V^{-1} - V^{-1}X'(X'V^{-1}X)^{-1}X'V^{-1}$, $W = ZZ'$, $G = 2[(y'Ay)y'AWA - (y'AWAy)y'A]$이며 B는 다음과 같다.

$$B = -\left(y'Ay\right)^2 \text{tr}\left(WV^{-1}WV^{-1}\right) - y'AWAyy'AWAy + 2y'(AWAWA)yy'Ay.$$

모의실험을 위하여 주어진 모형에서 먼저 균형 자료의 군집의 수 I와 균형 반복수 J를 정한다. 다음으로 총 관측치의 개수 $n = IJ$와 군집의 수 I는 고정하고 불균형 반복수 J_i를 변화시키면서 불균형성을 증가시켜서 균형 자료에서의 AIC의 효율과 비교한다. 고정 모수 $\mu = 0$으로 하고 오차항의 분산

표 **3.1.** 불균형 자료의 설계

조합	I	J	설계	n	군집에서의 반복수
	5		균형		3 3 3 3 3
1	5	3	불균형 I	15	2 2 3 4 4
	5		불균형 II		2 2 2 3 6
	5		균형		5 5 5 5 5
2	5	5	불균형 I	25	3 3 5 7 7
	5		균형 II		2 3 3 5 12
	7		균형		5 5 5 5 5 5 5
3	7	5	불균형 I	35	3 3 5 5 5 7 7
	7		균형 II		2 2 3 4 6 6 12
	5		균형		10 10 10 10 10
4	5	10	불균형 I	50	6 6 10 14 14
	5		균형 II		3 6 9 12 20
	10		균형		10 10 10 10 10 10 10 10 10 10
5	10	10	불균형 I	100	4 4 6 8 8 12 12 14 16 16
	10		균형 II		2 4 4 8 8 8 10 14 20 22

표 **3.2.** cAIC에서 모형의 복잡도 ρ의 값($\sigma^2 = 1$)

조합	설계	τ^2				
		0.1	0.2	0.4	0.6	0.8
	균형	1.92	2.50	3.18	3.57	3.82
1	불균형 I	1.90	2.45	3.12	3.51	3.76
	불균형 II	1.86	2.39	3.05	3.44	3.69
	균형	2.33	3.00	3.67	4.00	4.20
2	불균형 I	2.27	2.92	3.57	3.91	4.12
	균형 II	2.15	2.75	3.40	3.75	3.97
	균형	3.00	4.00	5.00	5.50	5.80
3	불균형 I	2.94	3.92	4.90	5.41	5.72
	균형 II	2.79	3.69	4.65	5.17	5.49
	균형	3.00	3.67	4.20	4.43	4.56
4	불균형 I	2.92	3.57	4.12	4.36	4.50
	균형 II	2.79	3.42	3.98	4.24	4.40
	균형	5.50	7.00	8.20	8.71	9.00
5	불균형 I	5.24	6.68	7.91	8.47	8.79
	균형 II	5.00	6.39	7.63	8.22	8.57

$\sigma^2 = 1$로 고정시키며 군집에 대한 분산성분은 6개를 고려하였다($\tau^2 = 0.0, 0.1, 0.2, 0.4, 0.6, 0.8$). 자료의 군집의 수 I와 반복수 J는 표 3.1에 나타나 있듯이 5개의 조합을 생각하고 각 조합에 대하여 1개의 균형 자료, 2개의 불균형 자료 I과 II를 고려하였다. 불균형 II 자료의 구조가 불균형 I 구조보다 불균형이 더 심하게 설계하였다. 유의할 점은 임의효과의 분산 τ^2이 0이면 상수평균모형(M_1)이 실제모형이고 임의효과의 분산 τ^2이 0보다 크면 임의절편모형(M_2)이 실제모형이다. 자료의 구성에 대한 조합은 표 3.1에 자세히 나타나 있다. 표 3.2는 식 (2.5)에 주어진 모형의 복잡성에 대한 측도인 ρ를 모든 모수를 알고 있다고 가정하고 그 참값을 계산한 것이다. ρ 값은 군집의 수 I가 많아질수록 또한 τ^2의 값이 커질수록 크게 된다. 반면 흥미로운 사실은 자료의 불균형성이 증가할수록 ρ 값은 작아진다.

표 **3.3.** mAIC를 사용할 때 모형 M_2 를 선택할 확률(1000번 모의실험, $\sigma^2 = 1$)

조합	설계	τ^2					
		0.0	0.1	0.2	0.4	0.6	0.8
1	균형	0.044	0.101	0.153	0.254	0.324	0.450
	불균형 I	0.048	0.071	0.151	0.228	0.341	0.405
	불균형 II	0.054	0.095	0.121	0.231	0.318	0.405
2	균형	0.030	0.118	0.221	0.434	0.551	0.651
	불균형 I	0.042	0.130	0.226	0.354	0.467	0.622
	균형 II	0.035	0.114	0.212	0.334	0.448	0.568
3	균형	0.029	0.148	0.328	0.560	0.712	0.798
	불균형 I	0.025	0.141	0.320	0.528	0.624	0.773
	균형 II	0.030	0.154	0.304	0.492	0.663	0.755
4	균형	0.045	0.242	0.435	0.715	0.782	0.866
	불균형 I	0.028	0.209	0.412	0.651	0.750	0.845
	균형 II	0.031	0.226	0.408	0.616	0.746	0.830
5	균형	0.048	0.422	0.726	0.909	0.976	0.991
	불균형 I	0.038	0.422	0.714	0.901	0.963	0.975
	균형 II	0.038	0.376	0.690	0.886	0.948	0.976

표 **3.4.** Vaida과 Blanchard (2005)의 cAIC를 사용할 때 모형 M_2 를 선택할 확률(1000번 모의실험, $\sigma^2 = 1$)

조합	설계	τ^2					
		0.0	0.1	0.2	0.4	0.6	0.8
1	균형	0.248	0.302	0.406	0.531	0.580	0.710
	불균형 I	0.218	0.303	0.391	0.498	0.601	0.660
	불균형 II	0.206	0.282	0.374	0.454	0.569	0.632
2	균형	0.243	0.406	0.540	0.716	0.806	0.848
	불균형 I	0.216	0.401	0.500	0.645	0.751	0.835
	균형 II	0.188	0.352	0.455	0.616	0.679	0.792
3	균형	0.224	0.470	0.647	0.785	0.880	0.932
	불균형 I	0.251	0.472	0.665	0.804	0.845	0.903
	균형 II	0.206	0.471	0.593	0.766	0.848	0.903
4	균형	0.228	0.600	0.744	0.886	0.921	0.955
	불균형 I	0.188	0.572	0.719	0.869	0.902	0.957
	균형 II	0.221	0.550	0.698	0.824	0.894	0.937
5	균형	0.330	0.772	0.918	0.980	0.998	0.997
	불균형 I	0.305	0.778	0.914	0.972	0.986	0.992
	균형 II	0.276	0.728	0.890	0.982	0.992	0.996

주어진 자료의 구조와 분산성분을 이용하여 관측치를 생성시키고 `R package`의 함수 `lme()`으로 모형을 적합시켰다. 적합된 모형에 대하여 (i) 식 (3.2)의 mAIC를 이용한 모형의 선택법, (ii) Vaida와 Blanchard (2005)에서 제안된 식 (2.6)의 보정항을 이용한 모형의 선택법, (iii) Greven과 Kneib (2010)에서 제안된 식 (3.4)의 보정항을 이용한 모형의 선택법을 각각 적용하여 복잡한 모형인 임의절편모형(M_2)을 선택하는 확률을 1000회 반복하여 추정하였다. 표 3.3은 mAIC를 사용했을 때 임의절편모형을 선택할 확률에 대한 모의실험의 결과이다. 복잡한 모형, 즉 실제 모형을 선택하는 확률은 분산성분 τ^2의 값이 커질수록 크게 되며 τ^2의 값이 고정되었을 때는 관측값의 수가 많아질수록 증가한다. 또한 불균형의 정도가 증가하면 자료가 균형인 경우보다 실제 모형을 선택하는 확률이 작아지는 경향을 발견할 수 있으나

표 **3.5.** Greven과 Kneib (2010)의 cAIC를 사용할 때 모형 M_2 를 선택할 확률(1000번 모의실험, $\sigma^2 = 1$)

조합	설계	τ^2					
		0.0	0.1	0.2	0.4	0.6	0.8
1	균형	0.188	0.234	0.326	0.446	0.500	0.638
	불균형 I	0.154	0.218	0.307	0.414	0.534	0.590
	불균형 II	0.132	0.200	0.260	0.378	0.495	0.552
2	균형	0.128	0.284	0.402	0.604	0.738	0.773
	불균형 I	0.139	0.268	0.363	0.511	0.642	0.766
	균형 II	0.104	0.234	0.350	0.509	0.589	0.711
3	균형	0.116	0.331	0.502	0.682	0.836	0.880
	불균형 I	0.140	0.300	0.516	0.694	0.776	0.858
	균형 II	0.106	0.314	0.438	0.652	0.788	0.864
4	균형	0.098	0.425	0.570	0.815	0.874	0.916
	불균형 I	0.100	0.382	0.565	0.774	0.848	0.914
	균형 II	0.102	0.398	0.568	0.748	0.846	0.896
5	균형	0.136	0.601	0.826	0.943	0.985	0.992
	불균형 I	0.136	0.612	0.822	0.940	0.982	0.991
	균형 II	0.102	0.520	0.808	0.958	0.974	0.990

그 차이는 크지 않다. mAIC를 이용한 모형선택의 결과는 우도비검정(likelihood ratio test)의 결과와 매우 밀접하며 실제 Akaike Information의 추정에서는 mAIC는 작은 모형을 선호하는 편이(bias)가 있다 (Greven과 Kneib, 2010).

표 3.4는 Vaida와 Blanchard (2005)의 cAIC를 사용했을 때 복잡한 모형을 선택할 확률에 대한 모의실험의 결과이다. mAIC를 이용할 경우와 비슷한 경향을 보이지만 복잡한 모형을 선택하는 확률이 mAIC에 비하여 매우 크다. 이는 분산성분 τ^2의 추정에 대한 불확실성을 반영하지 못한 결과이다. 표 3.5는 Greven과 Kneib (2010)의 cAIC를 사용했을 때 복잡한 모형을 선택할 확률에 대한 모의실험의 결과이다. mAIC와 cAIC를 이용할 경우와 비슷한 경향을 보이지만 복잡한 모형을 선택하는 확률이 Vaida와 Blanchard (2005)의 cAIC를 사용했을 경우보다 줄어들었다. 이는 분산성분 τ^2의 추정에 대한 불확실성을 반영하여 얻은 개선된 결과이다. 두 가지 cAIC를 사용한 결과도 불균형의 정도가 증가하면 자료가 균형인 경우보다 실제 모형을 선택하는 확률이 작아지는 경향을 발견할 수 있다.

4. 결론

본 논문은 불균형 자료에서 선형혼합모형에 적용되는 Akaike Information Criterion(AIC)의 효율에 대하여 알아보기 위하여 자료의 불균형이 모형선택 방법에 미치는 영향을 모의실험을 통하여 알아보았다. 자료의 불균형이 심해짐에 따라 AIC에 근거한 모든 모형선택방법이 복잡한 모형을 선택하는 확률이 작아지는 것을 알 수 있었다. 하지만 모의실험에서는 불균형과 균형의 차이가 크지 않다는 사실 또한 확인하였다. 또한 분산성분을 추정할 경우에 그 불확실성에 대한 정도를 cAIC에 반영한 Greven과 Kneib (2010)의 방법이 기존의 방법보다 합리적인 결과를 보여주는 것도 확인하였다. 모의실험에서 확인된 결과를 실제 자료 또는 더 복잡한 구조를 가지는 모형에 대하여 확장하는 것이 필요한 향후 연구 과제이다.

참고문헌

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Second International Symposium on Information Theory*, 267–281, Akademiai Kiado, Budapest.

Greven, S. and Kneib, T. (2010). On the behavior of marginal and conditional AIC in linear mixed models, *Biometrika*, **97**, 773–789.

Hartley, H. O. and Rao, J. N. K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model, *Biometrika*, **54**, 93–108.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association*, **72**, 320–338.

Jiang, J. (2009). *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, New York.

Khuri, A. I., Mathew, T. and Sinha, B. K. (1998). *Statistical Tests in Mixed linear Models*, John Wiley & Sons.

Liang, H., Wu, H. and Zou, G. (2008). A note on conditional AIC for linear mixed-effects models, *Biometrika*, **95**, 773–778.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components*, John Wiley & Sons.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models, *Biometrika*, **92**, 351–370.

Simulation Study on Model Selection Based on AIC under Unbalanced Design in Linear Mixed Effect Models

Yonghee Lee[1]

[1]Department of Statistics, University of Seoul

(Received September 2010; accepted October 2010)

Abstract

This article consider a performance model selection based on AIC under unbalanced deign in linear mixed effect models. Vaida and Balanchard (2005) proposed conditional AIC for model selection in linear mixed effect models when the prediction of random effects is of primary interest. Theoretical properties of cAIC and related criteria have been investigated by Liang *et al.* (2008) and Greven and Kneib (2010). However, all of the simulation studies were performed under a balanced design. Even though functional form of AIC remain same even under the unbalanced deign, it is worthwhile to investigate performance of AIC based model selection criteria under the unbalanced design. The simulation study in this article shows how unbalancedness affects model selection in linear mixed effect models.

Keywords: Linear mixed effect models, unbalanced design, AIC, model selection.

This Work was supported by the University of Seoul 2009 Research Fund.

[1]Professor, Department of Statistics, University of Seoul, 13 Siripdae-gil, Dongdaemun-gu, Seoul 130-743 Korea. E-mail: ylee@uos.ac.kr