The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.1.119
30(1), 119–133

Using the corrected Akaike's information criterion for model selection

Eunjung Song[a] · Sungho Won[b] · Woojoo Lee[a,1]

[a]Department of Statistics, Inha University;
[b]Department of Public Health Science, Seoul National University

(Received November 7, 2016; Revised December 14, 2016; Accepted January 7, 2017)

Abstract

Corrected Akaike's information criterion (AICc) is known to have better finite sample properties. However, Akaike's information criterion (AIC) is still widely used to select an optimal prediction model among several candidate models due to of a lack of research on benefits obtained using AICc. In this paper, we compare the performance of AIC and AICc through numerical simulations and confirm the advantage of using AICc. In addition, we also consider the performance of quasi Akaike's information criterion (QAIC) and the corrected quasi Akaike's information criterion (QAICc) for binomial and Poisson data under overdispersion phenomenon.

Keywords: AIC, AICc, QAIC, QAICc, overdispersion

1. 서론

통계분석에서 고려되는 여러 후보 모형 중에서 하나의 최적 모형을 선택하는 것은 매우 중요한 문제이지만 동시에 매우 어려운 문제이기도 하다. 이러한 모형선택 문제에 대한 하나의 해결책으로 Akaike (1973)는 Akaike's information criterion(AIC)를 제안하였다. AIC는 참모형과 후보 모형 사이의 불일치 정도를 수치화한 통계량으로, 각 모형에 대응하는 AIC 값 중 가장 작은 값에 대응하는 모형을 선택함으로써 예측력이 가장 좋은 모형을 선택하게 해주는 것으로 알려져 있다 (Shmueli, 2010). 그러나 AIC는 통계량을 유도하는 과정에서 최대가능도 추정량의 점근적 성질들을 이용하기 때문에 모형에서 추정해야 할 모수의 개수에 비해 자료수가 충분히 많지 않으면 예측력이 높은 모형을 선택하는데 어려움을 가진다. 이와 같은 한계점을 보완하기 위해 Hurvich와 Tsai (1989)는 corrected Akaike's information criterion(AICc)를 제안하였다. AICc는 정규분포를 가정한 회귀모형을 기반으로 AIC를 수정한 것으로, 유한표본 하에서 얻어지는 최대가능도 추정량의 정확한 분포(exact distribution)를 가지고 유도되는 통계량이다. 또한 AICc를 일반화 선형 모형으로 확장해서 사용할 수 있기 때문에, 많은 통계 모형의 선택 기준으로 기존의 AIC를 대체하여 AICc를 사용하는 것이 가능하다. 그러나 AICc의 계산이

This research was supported by a grant [MPSS-NH-2015-79] through the Disaster and Safety Management Institute funded by Ministry of Public Safety and Security of Korean government.
[1]Corresponding author: Department of Statistics, Inha University, 100, Inha-ro, Nam-gu, Incheon 22212, Korea. E-mail: lwj221@gmail.com

AIC를 계산하는 것에 비해 전혀 어렵지 않음에도 불구하고 다양한 분야의 연구에서 여전히 AIC만이 우선적으로 사용되고 있다 (Bloom과 Milkovich, 1998; Debrock 등, 2000; Harada 등, 2010; Johnson 등, 2016; Zampetakis 등, 2010). 따라서 AIC와 AICc의 성능을 시뮬레이션을 통해 비교해보고, AIC 대신 AICc를 우선적으로 고려함으로써 어떤 이득을 얻을 수 있는지 구체적으로 확인해 보는 것은 의미가 있다.

AIC와 AICc 비교연구는 Cavanaugh 등 (2008)에 의하여 시도된 적이 있었다. 이 연구에서 모형선택 기준들의 성능은 정규분포를 가정한 여러개의 후보모형 중에서 참모형이 얼마나 높은 빈도로 선택되었는지로 확인하였다. 그러나 AIC의 본질적 목표는 참모형을 찾는 것이 아니라, 예측을 잘하는 모형을 선택하기 위한 것이며 특히 예측을 잘하는 모형과 참모형은 동일하지 않은 경우가 빈번하다는 사실에 주목해야 한다 (Shmueli, 2010). 따라서 AIC와 AICc로 선택된 모형이 참모형인지 보다는 예측을 잘하는 모형인지를 확인하는 것이 바람직하다고 할 수 있으며, 본 논문에서 이러한 관점에서 선택된 모형을 평가하고자 한다. 그러나 선택된 모형이 예측력이 좋은 모형인지 아닌지를 확인하는 것은 참모형을 선택한 것인지 확인하는 것보다 조금 더 어려운 문제이다. 선택된 모형이 참모형인지의 여부를 단순히 확인했던 것과는 다르게 예측력이 좋은 모형인지 평가하기 위해서는 예측력을 평가하는 판단기준이 필요하기 때문이다. 본 논문에서는 예측력을 평가하는 판단기준으로 두 가지 측도를 이용하였다. 첫 번째는 피어슨 잔차의 제곱합을 이용하여 얻은 통계량이고, 두 번째 측도는 로그 가능도 함수를 이용하여 구한 통계량이다. 본 논문에서는 두 측도를 이용하여 예측력이 좋은 모형인지 아닌지 확인 할 뿐만 아니라, 두 측도를 비교해 봄으로써 선택된 모형의 예측력에 대한 평가가 사용되는 측도에 크게 의존하는지를 추가적으로 확인해 보고자 한다. 또한 Cavanaugh 등 (2008)의 연구에서는 참모형이 후보모형 중에 포함되어 있지 않은 경우는 다루어지지 않았는데 본 논문에서는 그에 대해서도 추가로 다루어 보고자 한다.

AIC는 정규분포를 가정한 선형 회귀 모형를 넘어서, 이산적인 형태의 자료 분석에서 널리 활용되고 있는 포아송 회귀 모형이나 로지스틱 회귀 모형에도 널리 활용되고 있다. 이 경우 AIC의 사용에 주의해야 할 점이 존재한다. 그것은 통계 모형에 반영되어 있지 않은 과대산포(overdispersion) 현상을 모형 선택에 어떻게 반영할 것인가이다. 여기서 과대산포란 일반화선형모형에서 포아송분포나 이항분포를 가정하는 경우 모형에 의해 기대되는 산포모수(dispersion parameter)보다 실제 추정된 산포 모수의 값이 더 큰 경우를 말한다. AIC와 AICc를 구성하는 로그가능도 함수 내부에 산포모수를 자연스럽게 반영하기가 어렵기 때문에, Lebreton 등 (1992)은 AIC와 AICc에 사용되는 로그가능도 함수를 추정된 산포모수로 나누어 준 새로운 모형선택기준을 제시하였고, 이를 각각 quasi Akaike's information criterion(QAIC)와 corrected quasi Akaike's information criterion(QAICc)라고 불렀다. QAIC와 QAICc의 성능 비교는 최근에 Kim 등 (2014)에 의해서 진행되었는데, 이 연구 또한 Cavanaugh 등 (2008)처럼 여러 후보모형 중에서 참모형이 선택되었는지 여부를 가지고 성능이 측정되었다. 그러므로 본 논문에서는 QAIC와 QAICc의 비교 연구에서도 예측력이 높은 모형을 얼마나 잘 선택 할 수 있는지를 가지고 성능을 확인하고자 한다.

본 논문의 구성은 다음과 같다. 제 2절에서 AIC와 AICc를 유도과정과 함께 소개한다. 제 3절에서는 정규분포를 이용한 시뮬레이션을 진행하여 AIC와 AICc가 예측력이 높은 모형을 얼마나 잘 선택하는지를 비교할 것이다. 제 4절에서는 QAIC와 QAICc를 소개하고, 포아송분포와 이항분포에서 시뮬레이션을 진행하여 두 기준의 성능을 비교해 볼 것이다. 제 5절에서 실제자료를 가지고 AIC, AICc, QAIC와 QAICc를 이용하여 모형을 선택하는 사례 연구를 진행할 것이고, 제 6절에서 결론을 내릴 것이다.

2. AIC와 AICc의 유도

통계 분석에서 모형을 선택하는 기준으로 많이 사용되는 AIC는 쿨백-라이블러 발산(Kullback-Leibler

divergence)을 기반으로 만들어진 통계량이다 (Akaike, 1973). 쿨백-라이블러 발산은 거리(distance)와 유사한 개념으로, 고려하고 있는 후보모형과 참모형 사이의 불일치 정도를 나타낸다. 참모형에 대응되는 결합확률분포가 $f(\boldsymbol{y}|\boldsymbol{\theta}^*)$이고 고려하고 있는 후보모형의 결합확률분포가 $g(\boldsymbol{y}|\boldsymbol{\theta})$인 경우 쿨백-라이블러 발산은

$$\Delta\left(f\left(\boldsymbol{y}|\boldsymbol{\theta}^*\right), g\left(\boldsymbol{y}|\theta\right)\right) = \int f\left(\boldsymbol{y}|\boldsymbol{\theta}^*\right) \log\left(\frac{f\left(\boldsymbol{y}|\boldsymbol{\theta}^*\right)}{g\left(\boldsymbol{y}|\boldsymbol{\theta}\right)}\right) d\boldsymbol{y} \tag{2.1}$$

으로 정의된다. 여기서 $\boldsymbol{y}$는 $n \times 1$ 랜덤 벡터이며 log는 자연로그를 의미한다. 위의 식 (2.1)에서 $\boldsymbol{\theta}$는 $K \times 1$ 차원 모수 벡터이고, 식 (2.1)을 최소로 만들어주는 $\boldsymbol{\theta}$를 $\boldsymbol{\theta_0}$로 나타낸다. 만약 후보모형에 참모형이 있는 경우에는, $\boldsymbol{\theta_0}$는 $\boldsymbol{\theta}^*$와 같게 되고, 그 때의 쿨백-라이블러 발산의 값은 0이 된다. 그 외의 모든 경우에는 쿨백-라이블러 발산이 항상 양수값을 가진다. 그러나 실제로 $\boldsymbol{\theta_0}$는 알려져 있지 않으므로 우리는 자료로 부터 모수 $\boldsymbol{\theta}$를 추정하여 식 (2.1)을 계산해야 한다. 이때 추정에 사용되는 자료 $\tilde{\boldsymbol{y}}$는 $\boldsymbol{y}$와는 독립이며 분포 $f(\cdot)$로 부터 추출된 랜덤표본이라고 가정한다. $\boldsymbol{\theta}$에 대한 추정량은 최대가능도 추정량(maximum likelihood estimator; MLE)을 이용하며 $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\tilde{\boldsymbol{y}})$으로 표기한다. $\widehat{\boldsymbol{\theta}}$은 자료 $\tilde{\boldsymbol{y}}$에 따라 값이 변하는 확률변수이므로 식 (2.1)에서 $\boldsymbol{\theta}$를 $\widehat{\boldsymbol{\theta}}$으로 대체한 후, 그것의 임의성을 제거하기 위해 기대값을 구해보면

$$E_{\widehat{\boldsymbol{\theta}}}\left[\Delta\left(f\left(\boldsymbol{y}|\boldsymbol{\theta}^*\right), g\left(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}\right)\right)\right] = E_{\boldsymbol{y}}\left[\log f\left(\boldsymbol{y}|\boldsymbol{\theta}^*\right)\right] - E_{\widehat{\boldsymbol{\theta}}}E_{\boldsymbol{y}}\left[\log g\left(\boldsymbol{y}|\widehat{\theta}\right)\right] \tag{2.2}$$

이 된다. 식 (2.2)는 항상 0 이상인 값을 가지고 $E_{\boldsymbol{y}}[\log f(\boldsymbol{y}|\boldsymbol{\theta}^*)]$는 상수항이므로, 식 (2.2)를 최소화하는 것은

$$T = E_{\widehat{\boldsymbol{\theta}}}E_{\boldsymbol{y}}\left[\log g\left(\boldsymbol{y}|\widehat{\theta}\right)\right] \tag{2.3}$$

를 최대화 하는 것과 같다. 부록 A을 참고하면, $-2T$에 대한 점근적인 불편추정량은

$$\mathrm{AIC} = -2\log g\left(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}\right) + 2K \tag{2.4}$$

임을 유도 할 수 있다. 위의 식 (2.4)에서 $-2\log g(\boldsymbol{y}|\widehat{\boldsymbol{\theta}})$은 적합도(goodness-of-fit)를 나타내는 항이라 해석 할 수 있다. 그러나 모수를 추정하여 $-2\log g(\boldsymbol{y}|\widehat{\boldsymbol{\theta}})$을 계산하기 때문에 $-2\log g(\boldsymbol{y}|\boldsymbol{\theta_0})$보다 더 작게 추정되는 편향이 나타나기 쉽다. 이러한 편향을 보정하기 위해 AIC에서는 $-2\log g(\boldsymbol{y}|\widehat{\boldsymbol{\theta}})$에 $2K$를 더해주는 것으로 이해 할 수도 있다. 만일 i번째 이진수자료(binary data) y_i의 성공확률 p_i에 대하여 로지스틱모형 $\log\left(p_i/(1-p_i)\right) = x_i^T\beta$을 사용한다면, 식 (2.4)에서

$$\log g\left(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}\right) = \log g\left(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}\right) = \sum\left\{y_i\left(x_i^T\widehat{\boldsymbol{\beta}}\right) - x_i^T\widehat{\boldsymbol{\beta}} - \log\left(1 + e^{-x_i^T\widehat{\boldsymbol{\beta}}}\right)\right\} \tag{2.5}$$

이며, K는 β의 개수와 같다. 또한 i번째 개수자료(count data) y_i의 기대값 μ_i에 대하여 포아송모형 $\log\mu_i = x_i^T\beta$를 사용하는 경우에는 식 (2.4)는

$$\log g\left(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}\right) = \log g\left(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}\right) = \sum\left\{y_i\left(x_i^T\widehat{\boldsymbol{\beta}}\right) - e^{x_i^T\widehat{\boldsymbol{\beta}}} - \log\left(y_i!\right)\right\} \tag{2.6}$$

이며, K는 로지스틱 회귀 모형과 마찬가지로 회귀계수의 개수가 된다.

AIC에서 편향을 보정하는 항인 $2K$는 후보모형들 가운데 참인 모형이 존재한다는 가정하에서 구해진다. 그러나 실제로 자료를 분석할 때 고려하고 있는 모형 가운데 참모형이 있을 가능성은 매우 낮다. 이와 같이 후보모형들 가운데 참모형이 없는 경우에 대하여 Takeuchi (1976)는 Takeuchi information criterion(TIC)를 제안하였으며, 이와 같은 가정의 완화는 TIC가 $2K$대신 $2\mathrm{tr}(\widehat{J}(\widehat{\boldsymbol{\theta}})[\widehat{I}(\widehat{\boldsymbol{\theta}})]^{-1})$ 을 패널

티 항으로 갖도록 한다. 여기서 $\widehat{J}(\widehat{\boldsymbol{\theta}})$은 $J(\boldsymbol{\theta_0}) = E_{\boldsymbol{y}}\left[[\partial/\partial\boldsymbol{\theta}\log g(\boldsymbol{y}|\boldsymbol{\theta})][\partial/\partial\boldsymbol{\theta}\log g(\boldsymbol{y}|\boldsymbol{\theta})]^{'}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}}$의 추정량이고 $\widehat{I}(\widehat{\boldsymbol{\theta}})$은 $I(\boldsymbol{\theta_0})$의 추정량이다. 그러나 실제자료를 분석할 때 TIC를 사용하는 경우는 거의 없는데, 이는 TIC의 편향을 보정하는 항이 자료에 따라 크게 바뀌는 등 불안정하기 때문이다 (Burnham과 Anderson, 2003).

AIC는 식 (2.3)을 테일러 근사하여 구한 추정량이지만 정규분포에서는 T를 정확하게 구할 수 있다. $n\times 1$벡터인 확률변수 $\boldsymbol{y}$가 정규분포를 따른다고 가정하면 $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, $\varepsilon \sim MVN(\boldsymbol{0}, \sigma^2 I_n)$으로 표현 할 수 있다. 이 식에서 I_n은 $n\times n$인 항등행렬을 나타내며, $\boldsymbol{X}$는 $n\times(K-1)$인 모형행렬(model matrix)이다. 또한 $\boldsymbol{\beta}$는 회귀계수를 의미하며 $(K-1)\times 1$ 벡터이고 σ^2은 오차항의 분산이다. 따라서 식 (2.1)의 θ는 $(\boldsymbol{\beta}, \sigma^2)$이다. 모수 $\boldsymbol{\beta}$와 σ^2의 최대가능도 추정량은 앞에서와 마찬가지로 $\boldsymbol{y}$와 독립이면서 같은 분포로 부터 추출된 랜덤표본 $\tilde{\boldsymbol{y}}$를 이용하여 구한다. 그리고 각각 $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\tilde{\boldsymbol{y}}) = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\tilde{\boldsymbol{y}}$, $\widehat{\sigma}^2 = \widehat{\sigma}^2(\tilde{\boldsymbol{y}}) = ((\tilde{\boldsymbol{y}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^{'}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}))/n$이라고 표현한다. 따라서 $\log g(\boldsymbol{y}|\widehat{\theta}) = \log g(\boldsymbol{y}|(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)) = -n/2\log\widehat{\sigma}^2 - 1/2[((\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^{'}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}))/\widehat{\sigma}^2]$으로 표현 할 수 있으므로 식 (2.3)은 다음과 같이 표현 할 수 있다.

$$T = E_{\widehat{\boldsymbol{\theta}}}E_{\boldsymbol{y}}\left[-\frac{n}{2}\log\widehat{\sigma}^2 - \frac{1}{2}\frac{\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^{'}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)}{\widehat{\sigma}^2}\right]. \tag{2.7}$$

부록 B를 따르면, $-2T$의 유한표본에서의 불편추정량은

$$\text{AICc} = -2\log g\left(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}\right) + \frac{2nK}{n-K-1} \tag{2.8}$$

가 된다. 또한 식 (2.8)의 $-2\log g(\boldsymbol{y}|\widehat{\boldsymbol{\theta}})$에 지수족의 로그가능도함수를 적용하여 일반화선형모형의 모형선택 문제에도 확장하여 사용할 수 있다. 즉, 로지스틱 모형과 포아송 모형에서는 $\log g(\boldsymbol{y}|\widehat{\boldsymbol{\theta}})$에 각각 식 (2.5)와 식 (2.6)를 이용하고, 식 (2.4)의 $2K$에 $n/(n-K-1)$을 곱한 보정항을 사용하여 AICc를 구할 수 있다. 여기서 AICc 역시 AIC처럼 후보모형 중에 참모형이 존재해야 한다는 가정을 가지고 있음에 유의해야 한다. 자료의 수 n이 충분히 커지면 $n/(n-K-1)$이 1에 가까워져 AIC와 AICc가 유사해진다.

다음의 3절에서는 지금까지 소개한 AIC와 AICc의 성능을 비교하고 확인하기 위하여 정규분포를 가정하여 시뮬레이션을 진행하였다. 실제 연구자들이 AIC를 사용할 때에는, 후보 모형이 참인 모형을 포함하는 여부에 관계없이 사용하는 경우가 많으므로 Cavanaugh 등 (2008)에서 고려된 시뮬레이션과는 달리 참모형이 후보모형들 가운데 없는 경우에 대해서도 수치 연구를 시행할 것이다.

3. 정규분포를 이용한 시뮬레이션

우리는 AIC와 AICc가 예측력이 높은 모형을 얼마나 잘 선택하는지 확인하기 위해 정규분포를 가정하여 시뮬레이션을 진행하였다. 시뮬레이션에서는 참모형으로부터 추출된 자료 $\tilde{y}_i$들을 이용하여 후보모형들의 모수 $\boldsymbol{\theta}$를 추정하여 $\widehat{\boldsymbol{\theta}}$을 구하고, 자료 $\tilde{y}_i$와는 독립이면서 같은 참모형으로부터 추출된 랜덤표본 y_i들을 이용하여 추정된 각각의 모형의 예측력을 평가하였다. 이 때, 예측성능을 판단하기 위한 기준으로 다음과 같이 두 개의 측도를 고려하였다. 첫 번째 측도는 피어슨 잔차의 제곱합인 피어슨의 χ^2 통계량이다.

$$\chi^2 = \sum\frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}, \tag{3.1}$$

여기서 $\mu_i = E(y_i)$를 나타내며 $\widehat{\mu}_i$은 그 추정값을 나타낸다. 그리고 분모에 있는 함수 $V(\mu_i)$는 일반화 선형 모형에서의 분산함수(variance funtion)를 의미한다. y_i가 모수 추론에 사용된 자료와 독립임을 가정할 때, 식 (3.1)의 값을 작게 하는 모형이 예측력이 높은 모형이라고 평가 할 수 있다.

두 번째 측도는 로그 가능도 함수 $g(\boldsymbol{y}|\boldsymbol{\theta})$를 이용한 것으로

$$-2\log g\left(\boldsymbol{y}|\widehat{\theta}\right) \tag{3.2}$$

이다. $g(\boldsymbol{y}|\widehat{\boldsymbol{\theta}})$은 일종의 적합도를 나타내므로 값이 커질 수록 예측력이 좋은 모형이라고 볼 수 있다. 그러나 식 (3.2)는 $g(\boldsymbol{y}|\widehat{\boldsymbol{\theta}})$에 -2를 곱했으므로 χ^2 통계량과 같이 값이 작을 수록 모형의 예측력이 좋다고 평가 할 수 있다.

시뮬레이션에서 반응변수는 $Y \sim N(\mu_0, \sigma^2)$에서 생성되었고, μ_0는

$$\mu_0 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 \tag{3.3}$$

이다. 이 모형이 참모형에 대응된다. σ^2은 4이고, 식 (3.3)에서 모든 회귀계수는 1이며 설명변수 X는 균일분포 Uniform(0, 10)으로부터 생성되었다.

먼저 참모형이 후보모형들 중에 포함되어 있는 경우를 고려하기 위해 X_1부터 X_{12}까지의 변수들을 모형에 차례대로 추가하여 12개의 후보 모형들을 구성하였다. 즉 첫 번째 모형은 절편 β_0와 X_1으로 구성된 선형모형이고 두 번째 모형은 첫 번째 모형에 X_2를 추가하여 얻은 모형이다. 이와 같은 방식으로 후보모형 가운데 여섯 번째 모형이 참모형이 되도록 하였다. 반면에, 참모형이 후보모형들 중에 포함되어 있지 않은 경우를 반영하기 위해 X_1, X_2, X_3와 X_7부터 X_{15}까지의 변수들을 이용하여 앞에서와 마찬가지로 12개의 후보모형을 구성하였다. 즉, 후보모형을 구성할 때에는 참모형에서 사용된 X_4, X_5, X_6을 사용하지 않았다. 자료수(n)는 50, 75, 100, 200으로 네 가지 경우가 고려되었으며 각각 1,000번씩 시뮬레이션이 진행되었다. 그리고 각 시뮬레이션마다 추가적으로 식 (3.3)으로부터 추출된 500개의 자료로 식 (3.1)과 식 (3.2)를 구하였으며, 그 값이 가장 작은 모형을 예측력이 가장 좋은 모형이라고 하였다.

AIC와 AICc에 의한 모형 선택 결과는 Table 3.1에 요약되어있다. Table 3.1 중에서 containing the true model은 참모형이 후보모형에 포함되어 있는 경우를 나타내며, not containing the true model은 참모형이 후보모형에 포함되어 있지 않은 경우를 나타낸다. Selection에 해당하는 부분은 success와 failure로 나눠져있다. Pearson's χ^2열의 success는 식 (3.1) 값이 가장 작은 모형과 AIC나 AICc로 선택한 모형이 일치함을 나타낸다. 이와 마찬가지로 loglikelihood 열은 식 (3.2) 값이 가장 작은 모형이 AIC나 AICc가 선택한 모형과 같은 경우를 나타낸다. 반면에 failure는 AIC와 AICc가 선택한 모형이 가장 작은 식 (3.1)의 값을 갖지 못하는 것을 의미한다. 따라서 success는 AIC나 AICc가 예측력이 가장 좋은 모형을 선택한 것을 나타내며 failure는 예측력이 가장 좋은 모형을 선택하지 못한 경우를 의미한다.

Table 3.1에서 보여주듯이 정규분포에서는 참모형을 포함하든 하지 않든 구분없이 AIC보다는 AICc가 상당히 더 좋은 성능을 가지고 있음을 알 수 있다. 이러한 경향은 자료의 개수나 예측력 판단의 기준에 상관없이 동일하게 나타나고 있다. 따라서, AIC 대신 AICc를 사용하는 것은 바람직해 보인다. Cavanaugh 등 (2008)에서는 같은 시뮬레이션 세팅에서 참모형을 선택한 결과를 요약했는데, 이 연구에서도 AIC보다는 AICc가 전반적으로 우수한 것으로 보고하였다. 이 두 결과로 부터, 우리는 정규분포로부터 얻어진 자료에서 AICc로 선택한 모형이 참모형이면서 예측력이 높은 모형일 가능성이 높다는 것을 유추해 볼 수 있다. 그러나 자료 수가 커질수록 AICc와 AIC의 성능은 서로 유사해진다는 것을 확인할 수 있었다.

Table 3.1. Model selection results by AIC and AICc for Gaussian linear models

n	Selection	Containing the true model				Not containing the true model			
		Pearson's χ^2		Loglikelihood		Pearson's χ^2		Loglikelihood	
		AIC	AICc	AIC	AICc	AIC	AICc	AIC	AICc
50	Success	407	573	445	622	397	534	422	558
	Failure	593	427	555	378	603	466	578	442
75	Success	403	517	438	558	450	525	461	536
	Failure	597	483	562	442	550	475	539	464
100	Success	420	492	430	508	425	482	439	497
	Failure	580	508	570	492	575	518	561	503
200	Success	411	448	420	458	373	405	374	407
	Failure	589	552	580	542	627	595	626	593

AIC = Akaike's information criterion; AICc = corrected Akaike's information criterion.

4. 과대산포를 고려한 AIC와 AICc

일반화선형모형에서 포아송분포나 이항분포를 가정한 일반화선형모형의 경우 산포모수(dispersion parmeter)는 1로 고정된다. 그러나 실제로 자료를 분석하는 경우 산포모수가 1보다 더 큰 경우가 발생하며, 이와 같은 경우 과대산포가 존재한다고 한다. 이 때, AIC나 AICc는 과대산포를 반영하여 모델을 선택하지 않기 때문에 예측력이 좋은 모형을 선택하기 어렵다. 이러한 한계점을 보완하기 위하여 Lebreton 등 (1992)은 다음의 두 가지 정보기준

$$\text{QAIC} = -2\frac{\log g\left(\boldsymbol{x}|\widehat{\theta}\right)}{\widehat{\phi}} + 2K, \tag{4.1}$$

$$\text{QAICc} = -2\frac{\log g\left(\boldsymbol{x}|\widehat{\theta}\right)}{\widehat{\phi}} + \frac{2nK}{n-K-1} \tag{4.2}$$

를 제안하였다. $\widehat{\phi}$은 추정된 산포모수를 의미하며 자유도 df로 피어슨의 χ^2 통계량을 나눠준

$$\widehat{\phi} = \frac{\chi^2}{\text{df}}$$

로 추정한다. 위의 식 (4.1)과 (4.2)에서 AIC와 AICc의 로그가능도 함수 $\log g(\boldsymbol{x}|\widehat{\theta})$ 대신에 $(\log g(\boldsymbol{x}|\widehat{\theta}))/\widehat{\phi}$을 사용함으로써 과대산포를 반영하도록 기존의 정보기준을 수정한 것을 확인 할 수 있다. 위의 두 식에서 K는 산포모수 ϕ를 포함한 모수의 수를 의미한다.

아래의 두 시뮬레이션은 포아송분포와 이항분포를 가정하여 시행한 것인데, 우리는 이 수치 연구의 결과로 부터 과대산포가 발생한 경우 QAIC와 QAICc의 성능을 확인하고, 기존의 AIC와 AICc와 비교해보고자 한다.

4.1. 포아송 분포를 이용한 시뮬레이션

y는 평균 μ_0에 대한 모형이

$$\log(\mu_0) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 \tag{4.3}$$

인 포아송분포에서 생성하였으며, 반응변수 X는 균일분포 Uniform(0, 10)으로부터 랜덤하게 추출되었다. 식 (4.3)에서 β_1, β_3, β_5는 -0.1이며 β_2, β_4, β_6는 0.1을 주었다. 그리고 β_0는 2.5로 두었으며,

Table 4.1. Model selection results by AIC, AICc, QAIC and QAICc: when the candidate Poisson models include the true model

n	Selection	Containing the true model							
		Pearson's χ^2				Loglikelihood			
		AIC	AICc	QAIC	QAICc	AIC	AICc	QAIC	QAICc
50	Success	464	579	448	574	458	580	446	575
	Failure	536	421	552	426	542	420	554	425
75	Success	439	527	444	524	430	523	433	515
	Failure	561	473	556	476	570	477	567	485
100	Success	442	502	433	507	438	499	432	504
	Failure	558	498	567	493	562	501	568	496
200	Success	401	419	403	417	405	422	407	416
	Failure	599	581	597	583	595	578	593	584

AIC = Akaike's information criterion; AICc = corrected Akaike's information criterion; QAIC = quasi Akaike's information criterion; QAICc = corrected quasi Akaike's information criterion.

Table 4.2. Model selection results by AIC, AICc, QAIC and QAICc: when the candidate Poisson models do not include the true model

n	Selection	Not containing the true model							
		Pearson's χ^2				Loglikelihood			
		AIC	AICc	QAIC	QAICc	AIC	AICc	QAIC	QAICc
50	Success	14	35	351	467	11	30	347	466
	Failure	986	965	649	533	989	970	653	534
75	Success	13	18	350	418	10	15	365	438
	Failure	987	982	650	582	990	985	635	562
100	Success	7	7	316	376	9	12	352	414
	Failure	993	993	684	624	991	988	648	586
200	Success	16	19	288	315	14	15	294	315
	Failure	984	981	712	685	986	985	706	685

AIC = Akaike's information criterion; AICc = corrected Akaike's information criterion; QAIC = quasi Akaike's information criterion; QAICc = corrected quasi Akaike's information criterion.

Tables 4.1과 4.2에 그 결과를 요약하였다. 각각의 표에서 containing the true model과 not containing the true model은 정규분포에서의 시뮬레이션처럼 후보모형 가운데 참모형이 포함되었는지 포함되어 있지 않은지를 표현하였다. 마찬가지로 success는 AIC, AICc, QAIC나 QAICc가 예측력이 가장 높은 모형을 선택한 것을 나타내고 failure는 예측력이 가장 좋은 모형을 선택하지 못한 경우를 의미하고 있다. 후보모형들은 제 3절에서 정규분포를 가정한 경우와 같은 방식으로 구성하였고 성능비교도 식 (3.1)과 (3.2)를 이용하였다. 여기서 주목할 만한 것은 후보모형이 참모형의 몇몇 변수를 포함하지 못하게 될 때 과대산포가 발생한다는 점이다.

참모형이 후보 모형에 포함되어 있는 Table 4.1의 경우에는 AICc가 AIC보다, QAICc는 QAIC보다 우수한 성능이 나타나고 있다. 그러나, AICc와 QAICc 사이에는 성능면에서 차이가 크게 나타나고 있지는 않았다. 그러나 참모형이 후보 모형에 포함되어 있지 않은 경우인 Table 4.2의 결과에서는 AICc의 성능이 매우 떨어지는 것으로 나타나었고, 기본적으로 QAICc가 사용되어야 함이 잘 나타나고 있다. 이 결과는 과대산포를 고려한 모형기준을 사용해야 함을 나타내준 결과라고 볼 수 있다. 또한 QAICc가 QAIC보다 유한표본에서 더 나은 예측력을 가지고 있는 모형을 더 정확하게 선택하는 경향이 있음을 확

Table 4.3. Model selection results by AIC, AICc, QAIC and QAICc: when the candidate binomial models include the true model ($m = 5$)

n	Selection	Containing the true model							
		Pearson's χ^2				Loglikelihood			
		AIC	AICc	QAIC	QAICc	AIC	AICc	QAIC	QAICc
50	Success	419	521	405	523	471	596	465	599
	Failure	581	479	595	477	529	404	535	401
75	Success	410	498	417	499	442	535	451	538
	Failure	590	502	583	501	558	465	549	462
100	Success	434	495	438	498	470	528	467	536
	Failure	566	505	562	502	530	472	533	464
200	Success	396	422	395	420	408	435	407	431
	Failure	604	578	605	580	592	565	593	569

AIC = Akaike's information criterion; AICc = corrected Akaike's information criterion; QAIC = quasi Akaike's information criterion; QAICc = corrected quasi Akaike's information criterion.

인 할 수 있었다. 이와 유사한 결과가 Kim 등 (2014)에서도 언급되었는데, 참모형을 QAIC, QAICc 가 기존의 AIC, AICc보다 잘 선택한다는 것이다. 따라서 QAICc로 선택한 모형이 참모형이면서 예측력이 높은 모형일 가능성이 높다고 생각할 수 있다.

4.2. 이항분포를 이용한 시뮬레이션

y는 이항분포 binomial(m, p_0)에서 생성하고, p_0에 대한 모형은

$$\text{logit}(p_0) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 \tag{4.4}$$

으로 주어진다. m은 시행 수(binomial denominator)를 의미하며 시뮬레이션에서는 각각 5와 50이 고려되었다. 이항분포를 가정한 경우에도 앞에서 고려한 두 확률분포와 마찬가지로 균일분포 Unifrom(0, 10)으로부터 설명변수 X를 추출하였다. 그리고 회귀계수 β_0, β_1, β_3, β_5는 각각 0.3이며 β_2, β_4, β_6는 -0.3이다. 식 (4.4)에서 $m = 5$인 경우에는 Tables 4.3과 4.4에 결과를 요약했으며, $m = 50$인 경우는 Tables 4.5와 4.6에 결과를 정리했다.

이항분포의 결과도 포아송분포와 유사하게 참모형이 존재하는 경우에는 AICc와 QAICc가 AIC와 QAIC보다 예측력이 좋은 모형을 더 많이 선택하는 경향을 보여주며, 둘 사이의 성능은 비슷한 것으로 나타났다. 그러나 참모형이 후보모형에 존재하지 않으면 QAICc의 성능이 가장 좋게 나타나는 경향이 두드러지며, 이러한 경향성은 m이 5일 때보다 50일 때 더 두드러지게 나타나는 것을 Tables 4.5와 4.6을 통해 확인 할 수 있다.

5. 실제자료

지금까지 우리는 AIC, AICc, QAIC와 QAICc에 대하여 알아보고 시뮬레이션을 통하여 그들 간의 성능을 비교하여 보았다. 이를 토대로 제 5절에서는 위의 네 가지 기준들을 실제자료에 적용하여 보고자 한다. 실제 많은 연구에서 여전히 AIC와 QAIC가 많이 사용되고 있지만, 수치연구 결과에서 살펴보았듯이 AICc와 QAICc가 예측력이 좋은 모형을 선택하는데 더 우수한 성능을 보여주었고 실제 자료를 다룰 때에 AIC와 QAIC와는 다른 모형이 AICc와 QAICc에 의해 선택되어짐을 보여주고자 한다.

Table 4.4. Model selection results by AIC, AICc, QAIC and QAICc: when the candidate binomial models do not include the true model ($m = 5$)

n	Selection	Not containing the true model							
		Pearson's χ^2				Loglikelihood			
		AIC	AICc	QAIC	QAICc	AIC	AICc	QAIC	QAICc
50	Success	96	193	344	388	132	254	455	532
	Failure	904	807	656	612	868	746	545	468
75	Success	124	175	385	451	143	207	452	538
	Failure	876	825	615	549	857	793	548	462
100	Success	134	192	427	474	137	200	454	502
	Failure	866	808	573	526	863	800	546	498
200	Success	125	140	393	417	114	134	395	415
	Failure	875	860	607	583	886	866	605	585

AIC = Akaike's information criterion; AICc = corrected Akaike's information criterion; QAIC = quasi Akaike's information criterion; QAICc = corrected quasi Akaike's information criterion.

Table 4.5. Model selection results by AIC, AICc, QAIC and QAICc: when the candidate binomial models include the true model ($m = 50$)

n	Selection	Containing the true model							
		Pearson's χ^2				Loglikelihood			
		AIC	AICc	QAIC	QAICc	AIC	AICc	QAIC	QAICc
50	Success	474	577	471	585	482	586	477	592
	Failure	526	423	529	415	518	414	523	408
75	Success	454	535	456	531	452	533	449	527
	Failure	546	465	544	469	548	467	551	473
100	Success	472	540	466	542	480	549	476	550
	Failure	528	460	534	458	520	451	524	450
200	Success	401	433	395	425	404	436	398	427
	Failure	599	567	605	575	596	564	602	573

AIC = Akaike's information criterion; AICc = corrected Akaike's information criterion; QAIC = quasi Akaike's information criterion; QAICc = corrected quasi Akaike's information criterion.

Table 4.6. Model selection results by AIC, AICc, QAIC and QAICc: when the candidate binomial models do not include the true model ($m = 50$)

n	Selection	Not containing the true model							
		Pearson's χ^2				Loglikelihood			
		AIC	AICc	QAIC	QAICc	AIC	AICc	QAIC	QAICc
50	Success	0	1	414	508	0	1	443	551
	Failure	1000	999	586	492	1000	999	557	449
75	Success	0	2	451	531	0	1	458	534
	Failure	1000	998	549	469	1000	999	542	466
100	Success	0	0	456	510	0	0	444	496
	Failure	1000	1000	544	490	1000	1000	556	504
200	Success	4	4	390	414	6	6	366	395
	Failure	996	996	610	586	994	994	634	605

AIC = Akaike's information criterion; AICc = corrected Akaike's information criterion; QAIC = quasi Akaike's information criterion; QAICc = corrected quasi Akaike's information criterion.

Table 5.1. Variables description of air pollution data and estimated regression coefficients (standard error)

변수	설명	AIC로 선택한 모형	AICc로 선택한 모형
X_1	연평균 강우량	1.649 (0.603)	1.797 (0.599)
X_2	1월 평균온도	−1.893 (0.589)	−1.484 (0.514)
X_3	7월 평균온도	−2.300 (1.234)	−2.355 (1.244)
X_4	65세 이상인구비율		
X_5	가구당 인구수	−62.017 (44.782)	
X_6	교육기간의 중앙값	−16.966 (6.820)	−13.619 (6.432)
X_7	건전한 가구의 비율		
X_8	제곱마일당 인구수		
X_9	유색인종의 비율	5.216 (0.827)	4.585 (0.696)
X_{10}	사무직 고용의 비율		
X_{11}	소득 $3,000이하의 비율		
X_{12}	탄화수소의 상대적오염잠재력		
X_{13}	질산화물의 상대적오염잠재력		
X_{14}	이산화유황의 상대적오염잠재력	0.225 (0.082)	0.260 (0.078)
X_{15}	상대습도		
Y	연령의 의한 효과가 조정된 사망률		

AIC = Akaike's information criterion; AICc = corrected Akaike's information criterion.

5.1. 대기오염자료

대기오염에 대한 자료는 McDonald와 Schwing (1973)에서 연구되었던 자료이며, 15개의 설명변수와 1개의 반응변수로 구성되어 있다. 반응변수 Y는 연령의 효과가 조정된 사망률이며, 설명변수는 기후와 관련된 변수와 사회경제적 요인들에 대한 변수들로 구성되어있다. Table 5.1에는 이러한 반응변수와 설명 변수들에 대한 설명이 요약되어 있으며, 자료의 수는 60개이다.

대기오염 자료에 대한 분석은 정규분포를 가정한 회귀분석으로 진행하였으며, 1차항만을 가진 모든 조합의 모형들 중에서 AIC와 AICc를 이용하여 최종모형을 선택하였다. 총 32,767개의 모형 중에서 AIC가 가장 작은 모형은 $X_1, X_2, X_3, X_5, X_6, X_9, X_{14}$를 포함하고 있으며 AIC는 601.103이고, AICc는 604.703이다. 한편 AICc가 가장 작은 모형은 위의 모형에서 가구당 인구수를 나타내는 X_5가 제외된 모형이 선택되었으며, 이 모형의 AICc는 604.099이고, AIC는 601.276이었다. 위의 두 모형의 회귀계수 추정값과 그 표준오차는 Table 5.1에 함께 요약되어 있다. 두 모형의 AICc 값의 차이가 작기 때문에 실제 후자의 모형이 예측력의 관점에서 확실하게 우수하다기 보다는, 랜덤하게 더 작은 AICc 값을 가지게 되었을 가능성을 배제할 수는 없다. 따라서 추가로 두 모형의 예측력을 평가하기 위해 leave one out cross validation(LOOCV)를 이용하여 식 (3.1)과 (3.2)를 계산하였다. AIC로 선택된 모형은 식 (3.1)과 (3.2)의 값이 각각 79,645.400과 609.916으로 나타났다. 그리고 AICc를 이용하여 선택된 모형에서는 각각 79,492.460과 609.150이었다. 두 값들을 각각 비교하였을 때, AICc가 선택한 모형이 예측력이 더 좋을 것이라는 기대를 할 수 있다. 이처럼 AIC와 AICc가 선택하는 모형은 일치하지 않을 수 있기 때문에 분석자는 AIC를 항상 우선적으로 사용하는 분석 습관을 지양하고 AICc를 함께 고려하여 신중하게 모형을 선택할 필요가 있다.

5.2. 발아실험자료

앞에서 소개된 5.1절과 마찬가지로 실제 자료를 분석 할 때, QAICc와 QAIC가 선택하는 모형은 다를 수 있다. 이를 확인하기 위해 사용된 자료는 Hinde와 Demetrio (2007)에 소개되어 있는 자료이며, 초

Table 5.2. Variables description of orobanche seed germination data and estimated regression coefficients (standard error)

변수	설명	QAIC로 선택한 모형	QAICc로 선택한 모형
X_1	뿌리추출객의 종류(콩 = 0, 오이 = 1)	0.540 (0.250)	1.057 (0.144)
X_2	씨앗의 종류(O. aegyptiaca 73 = 0, O. aegyptiaca 75 = 1)	−0.146 (0.223)	
X_1X_2	교차항	0.778 (0.306)	
Y	실험된 씨앗중에서 발아한 씨앗의 개수		

QAIC = quasi Akaike's information criterion; QAICc = corrected quasi Akaike's information criterion.

종용(orobanche)이라는 식물에 속하는 두 종류의 씨앗이 오이뿌리 추출물과 콩뿌리 추출물에서 발아하는 것을 실험한 자료이다. 반응변수는 m_i개의 씨앗 중에서 y_i개의 씨앗이 발아되었는지에 대한 것이며, y_i를 이진수자료 관점에서 보았을 때 자료의 수는 총 831개이다. 그리고 설명변수는 실험에 사용된 씨앗 두 가지 종류와 각 씨앗이 어떤 추출물에서 발아되었는지에 대한 것이다 (Table 5.2). 따라서 두 개의 설명변수에 대하여 반응변수는 이항분포를 따른다고 가정하여 일반화선형모형으로 분석하였다. 고려한 후보모형은 각 변수가 하나씩 들어간 것과 두 개 모두 포함된것, 교차항이 포함된 것까지 총 4개를 고려하였다. 따라서 발아될 확률 p에 대하여

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2 \tag{5.1}$$

이 가장 복잡한 모형이다. AIC와 AICc는 식 (5.1)의 모형에서 가장 작은 값을 갖고, 그 값은 각각 117.874와 120.374이다. 그러나 식 (5.1)에서 추정된 산포모수의 값은 1.862로 이론적인 산포모수 값인 1에 비하여 거의 2배에 가깝다. 앞의 시뮬레이션으로 확인해 본 것처럼, 과대산포가 의심되어지는 경우에는 QAIC와 QAICc를 이용하는 것이 예측력이 높은 모형을 선택할 가능성이 더 높다. 발아실험 자료에서 고려된 모형 중에서 QAIC가 가장 작은 값을 가진 모형은 교차항까지 포함된 식 (5.1)이며, 이 모형에서 QAIC와 QAICc는 각각 69.014와 73.014이다. 한편 QAICc가 선택한 모형은 X_1만 포함된 모형으로, QAIC는 70.102이고 QAICc는 71.514이다. 즉 AIC, AICc와 QAIC가 선택한 모형은 식 (5.1)로 동일하며, QAICc으로 선택된 모형은 X_1만 포함한 모형이다. 이 두 모형의 회귀계수 추정값과 표준오차는 Table 5.2에 요약되어 있다. 대기오염자료에서와 마찬가지로 두 모형의 예측력을 좀 더 자세히 비교하기 위해 LOOCV를 통하여 식 (3.1)과 (3.2)를 계산하였다. 먼저 AIC, AICc, QAIC에 의해 선택된 모형은 식 (3.1)과 (3.2)의 값이 각각 45,922.660, 130.236으로 나타났다. 그리고 QAICc를 이용하여 선택한 모형에서는 각각 44,768.940과 129.246이였다. 따라서, 본 자료에서는 QAICc에 의해 선택된 모형의 예측력이 좀 더 우수한 것으로 판단된다. 그러나, QAIC에 의해 선택된 모형과는 해석 상 큰 차이가 있기 때문에 어떤 모형을 사용하는 것이 좋은지에 대해 전문가와의 토론 등의 추가적인 검토가 필요해 보인다.

6. 결론

우리는 시뮬레이션 연구를 통하여 AIC와 AICc의 성능을 비교하였다. 이를 통해 후보모형에 참모형이 포함되어있는지 여부와 무관하게 AIC보다 AICc가 예측력이 높은 모형을 더 잘 선택하는 것으로 나타났다. 또한 실제자료를 분석함으로써 AIC나 AICc가 같은 모형을 선택하지 않는 경우가 있을 수 있음을 확인하였다. 따라서 우리는 여러 후보 모형 중에서 하나의 모형을 선택할 때, AIC만을 이용하는 현재의 습관대신에 AICc를 이용하여 신중하게 모형을 선택하는 것이 바람직해 보인다. 또한 과대산포가 존재하는 경우에는 AIC나 AICc보다는 QAIC나 QAICc를 이용하는 것이 예측력이 높은 모형

을 선택할 가능성을 높여준다는 것을 보여주었다. 특히 QAICc의 성능이 가장 우수하였으므로, 이항분포나 포아송분포를 가정하여 자료를 분석할 때 과대산포가 의심된다면 QAICc를 사용하여 모형선택 하는 것이 바람직하다. 앞으로의 연구에서는 AICc와 QAICc의 성능이 일반화 선형 모형의 연결함수(link function)에 따라 어떠한 영향을 받는지 자세하게 살펴볼 필요가 있다고 생각된다.

부록 A

먼저 식 (2.3)에서 $E_{\boldsymbol{y}}[\log g(\boldsymbol{y}|\widehat{\theta})]$의 근사식을 구한다. $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta_0}$에서 테일러 전개를 한 다음, $f(\boldsymbol{y}|\boldsymbol{\theta^*})$에 대하여 기대값을 구하면

$$E_{\boldsymbol{y}}\left[\log g\left(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}\right)\right] \approx E_{\boldsymbol{y}}[\log g(\boldsymbol{y}|\boldsymbol{\theta_0})] + \left[E_{\boldsymbol{y}}\left[\frac{\partial \log g\left(\boldsymbol{y}|\theta\right)}{\partial \theta}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]'\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0}\right) - \frac{1}{2}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0}\right)' I\left(\boldsymbol{\theta_0}\right)\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0}\right)$$

$$= E_{\boldsymbol{y}}\left[\log g\left(\boldsymbol{y}|\boldsymbol{\theta_0}\right)\right] - \frac{1}{2}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0}\right)' I\left(\boldsymbol{\theta_0}\right)\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0}\right) \tag{A.1}$$

를 얻을 수 있으며, $I(\boldsymbol{\theta_0}) = E_{\boldsymbol{y}}[-[\partial^2 \log\left(g(\boldsymbol{y}|\boldsymbol{\theta})\right)]/\partial\theta\partial\theta']_{\boldsymbol{\theta}=\boldsymbol{\theta_0}}$이다. $\boldsymbol{\theta_0}$는 $\Delta(f(\boldsymbol{y}|\boldsymbol{\theta^*}), g(\boldsymbol{y}|\boldsymbol{\theta}))$를 최소화하는 값이므로, 식 (A.1)에서 $E_{\boldsymbol{y}}[(\partial \log g(\boldsymbol{y}|\boldsymbol{\theta}))/\partial\boldsymbol{\theta}]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}}$는 0이 된다.

그리고 식 (A.1)를 식 (2.3)에 대입하여 다음의 식을 구할 수 있다.

$$T \approx E_{\boldsymbol{y}}\left[\log g\left(\boldsymbol{y}|\boldsymbol{\theta_0}\right)\right] - \frac{1}{2}\mathrm{tr}\left\{I\left(\boldsymbol{\theta_0}\right)\Sigma\right\}. \tag{A.2}$$

위의 식 (A.2)에서 $\Sigma = E_{\widehat{\boldsymbol{\theta}}}\{[\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}][\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}]'\}$ 이며, $\mathrm{tr}(\cdot)$는 행렬의 대각합을 의미한다. 그 다음으로, 최대가능도 추정량 $\widehat{\theta} = \widehat{\theta}(\boldsymbol{y})$를 이용하여 식 (A.2)의 $E_{\boldsymbol{y}}[\log g(\boldsymbol{y}|\boldsymbol{\theta_0})]$를 근사한다. 먼저 $\boldsymbol{\theta_0} = \widehat{\boldsymbol{\theta}}(\boldsymbol{y})$에서 테일러 전개하고 $f(\boldsymbol{y}|\theta^*)$에 대하여 기대값을 구하면

$$E_{\boldsymbol{y}}\left[\log g\left(\boldsymbol{y}|\boldsymbol{\theta_0}\right)\right] \approx E_{\boldsymbol{y}}\left[\log g\left(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}\right)\right] - \frac{1}{2}\mathrm{tr}\{I\left(\boldsymbol{\theta_0}\right)\Sigma\} \tag{A.3}$$

을 얻을 수 있다. 이를 이용하면

$$T \approx E_{\boldsymbol{y}}\left[\log g\left(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}\right)\right] - \mathrm{tr}\{I\left(\boldsymbol{\theta_0}\right)\Sigma\}$$

이 된다. 그리고 만일 고려되는 후보모형 중에 참모형이 포함되어 있으면, $I(\boldsymbol{\theta_0}) = \Sigma^{-1}$이 되어서 $\mathrm{tr}\{I(\boldsymbol{\theta_0})\Sigma\} = K$가 된다. 따라서 T의 추정량으로

$$\widehat{T} \approx \log g\left(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}\right) - K \tag{A.4}$$

을 사용할 수 있다. 그리고 위의 식 (A.4)에 -2를 곱하면 다음과 같이 AIC를 얻을 수 있다.

$$\mathrm{AIC} = -2\log g\left(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}\right) + 2K. \tag{A.5}$$

부록 B

유도과정의 단순화를 위하여 $f(\boldsymbol{y}|\boldsymbol{\theta^*})$와 $g(\boldsymbol{y}|\boldsymbol{\theta_0})$가 동일하다고 가정하면 $E_{\boldsymbol{y}}[(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})] = n\sigma^2 + (\boldsymbol{X}\beta - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{X}\beta - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$을 얻을 수 있다. 그리고 이것을 식 (2.7)에 적용하여 다음을 얻었다.

$$E_{\widehat{\theta}}\left[-\frac{n}{2}\log\widehat{\sigma}^2\right] - \frac{1}{2}E_{\widehat{\theta}}\left[\frac{n\sigma^2 + \left(\boldsymbol{X}\beta - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}\beta - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)}{\widehat{\sigma}^2}\right]. \tag{B.1}$$

식 (B.1)의 두 번째 항에서 $\widehat{\boldsymbol{\beta}}$과 $\widehat{\sigma}^2$는 독립이므로 분리해서 계산을 할 수 있다. 그리고 추정량 $\widehat{\boldsymbol{\beta}}$의 공분산행렬이 $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$이므로 $E_{\widehat{\boldsymbol{\theta}}}[(\boldsymbol{X}\beta-\boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{X}\beta-\boldsymbol{X}\widehat{\boldsymbol{\beta}})] = \sigma^2(K-1)$이 된다. 이 결과를 식 (B.1)에 적용해보면

$$E_{\widehat{\boldsymbol{\theta}}}\left[-\frac{n}{2}\log\widehat{\sigma}^2\right] - \frac{\sigma^2}{2}(n+K-1)E_{\widehat{\boldsymbol{\theta}}}\left[\frac{1}{\widehat{\sigma}^2}\right]$$

이 된다. 위의 식에 $n\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{n-K+1}$을 적용하고, 자유도 df를 가진 카이제곱분포 χ^2_{df}에 대한 $1/\chi^2_{\mathrm{df}}$의 기대값이 $1/(\mathrm{df}-2)$인 사실을 이용하면

$$T = E_{\widehat{\boldsymbol{\theta}}}\left[-\frac{n}{2}\log\widehat{\sigma}^2 - \frac{n}{2}\right] - \frac{nK}{n-K-1} \tag{B.2}$$

를 얻을 수 있다. 위의 식 (B.2)은 후보모형 $g(\boldsymbol{y}|\theta)$가 참모형 $f(\boldsymbol{y}|\theta^*)$를 포함하고 있을 때만 성립하며, T에 대한 추정량으로

$$\widehat{T}^* = -\frac{n}{2}\log\widehat{\sigma}^2 - \frac{n}{2} - \frac{nK}{n-K-1} \tag{B.3}$$

을 얻을 수 있다. 식 (B.3)의 첫 번째 항과 두 번째 항은 $\widehat{\theta} = \widehat{\theta}(\boldsymbol{y})$를 이용한 정규분포의 $\log g(\boldsymbol{y}|\widehat{\boldsymbol{\theta}})$임을 알 수 있다. 이를 이용하여 AICc는 $(-n/2)\log\widehat{\sigma}^2 - n/2$ 대신에 로그 가능도 함수 $\log g(\boldsymbol{y}|\widehat{\boldsymbol{\theta}})$를 대체하고 -2를 곱한

$$\mathrm{AICc} = -2\log g\left(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}\right) + \frac{2nK}{n-K-1} \tag{B.4}$$

로 정의된다.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (pp. 267–281), Akadémia Kiadó, Budapest.

Bloom, M. and Milkovich, G. T. (1998). Relationships among risk, incentive pay, and organizational performance, *Academy of Management Journal*, **41**, 283–297.

Burnham, K. P. and Anderson, D. (2003). *Model Selection and Multi-Model Inference: a Practical Informatio-Theoric Approach*, Springer, New York.

Cavanaugh, J. E., Davies S. L., and Neath, A. A. (2008). Discrepancy-based model selection criteria using cross-validation. In *Statistical Models and Methods for Biomedical and Technical Systems* (pp. 473–486), Birkhauser, Boston.

Debrock, C., Preux, P. M., Houinato, D., Druet-Cabanac, M., Kassa, F., Adjien, C., Avode, G., Denis, F., Boutros-Toni, F., and Dumas, M. (2000). Estimation of the prevalence of epilepsy in the Benin region of Zinvié using the capture-recapture method, *International Journal of Epidemiology*, **29**, 330–335.

Harada, T., Ariyoshi, N., Shimura, H., Sato, Y., Yokoyama, I., Takahashi, K., Yamagata, S., Imamaki, M., Kobayashi, Y., Ishii, I., Miyazaki, M., and Kitada, M. (2010). Application of Akaike information criterion to evaluate warfarin dosing algorithm, *Thrombosis Research*, **126**, 183–190.

Hinde, J. and Demetrio, C. G. B. (2007). Overdispersion: models and estimation. In *A Short Course for 13th Brazilian Symposium of Probability and Statistics (SINAPE 1998)*, Brazil.

Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples, *Biometrika*, **76**, 297–307.

Johnson, R. J., Kerr, C. L., Enouri, S. S., Modi, P., Lascelles, B. D. X., and Castillo, J. R. E. (2016). Pharmacokinetics of liposomal encapsulated buprenorphine suspension following subcutaneous administration to cats, *Journal of Veterinary Pharmacology and Therapeutics*, Available from: http://dx.doi.org/10.1111/jvp.12357

Kim, H. J., Cavanaugh, J. E., Dallas, T. A., and Fore, S. A. (2014). Model selection criteria for overdispersed data and their application to the characterization of a host-parasite relationship, *Environmental and Ecological Statistics*, **21**, 329–350.

Lebreton, J. D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies, *Ecological Monograph*, **62**, 67–118.

McDonald, G. C. and Schwing, R. C. (1973). Instabilities of regression estimates relating air pollution to mortality, *Technometrics*, **15**, 463–481.

Shmueli, G. (2010). To explain or to predict?, *Statistical Science*, **25**, 289–310.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting, *Suri-Kagaku (Mathematic Sciences)*, **153**, 12–18.

Zampetakis, L. A., Bouranta, N., and Moustakis, V. S. (2010). On the relationship between individual creativity and time management, *Thinking Skills and Creativity*, **5**, 23–32.

모형 선택에서의 수정된 AIC 사용에 대하여

송은정[a] · 원성호[b] · 이우주[a,1]

[a]인하대학교 통계학과, [b]서울대학교 보건대학원

(2016년 11월 7일 접수, 2016년 12월 14일 수정, 2017년 1월 7일 채택)

Abstract

이미 corrected Akaike's information criterion(AICc)가 AIC에 비해 우수한 이론적 성질을 가진 것으로 알려져 있으나, 현재 실제 자료분석에서 최적의 예측 모형을 선택하기 위해 가장 널리 사용되는 정보기준은 여전히 Akaike's information criterion(AIC)이다. 이것은 AICc를 사용함으로써 실제 우리가 어떠한 종류의 이점을 얻을 수 있는가에 대해 논의하고 있는 연구가 부족해서이다. 우리는 이 논문에서 수치 연구를 통해 AIC와 AICc의 성능을 비교하고 AICc 의 사용이 가져오는 장점에 대해 확인을 할 것이다. 또한, 포아송 또는 이항 분포 자료 분석에서 과대산포(overdispersion) 현상이 나타난 경우 사용하는 quasi Akaike's information criterion(QAIC)와 corrected quasi Akaike's information criterion(QAICc) 성능에 대해서도 시뮬레이션을 통해 비교해보고자 한다.

주요용어: AIC, AICc, QAIC, QAICc, 과대산포

본 연구는 정부(국민안전처)의 재원으로 재난안전기술개발사업단의 지원을 받아 수행된 연구임[MPSS-자연-2015-79].

[1]교신저자: (22212) 인천광역시 남구 인하로 100, 인하대학교 통계학과. E-mail: lwj221@gmail.com