Akaike's criteria

Jieun Shin

2022-09-13

이 글은 AIC와 BIC의 의미와 유도과정을 알아보고자 한다.

AIC

참 모델을 $g(y)$, 후보 모델을 $f(y|\beta_j) \in \mathcal{F}$ 그리고 적합된 모델을 $f(y_i|\hat{\beta}_j)$이라고 하자. AIC는 기본적으로 적합된 모델과 참 모델 사이의 거리를 측정하는 방법으로, AIC를 가장 작게 하는 적합된 모델을 가장 좋은 모델로 여긴다. 이 때 측정은 K-L information으로 한다.

참 모델 $g(y_i)$과 적합된 모델 $f(y_i|\hat{\beta}_j)$ 사이의 K-L information은 다음과 같이 나타낸다:

$$I(\beta_j) = \mathbb{E}\left[\log \frac{g(y)}{f(y_i|\hat{\beta}_j)}\right]$$

여기서 $\mathbb{E}$는 $g(y)$ 하에서의 기댓값이다. 그러면 Kullback discrepancy (불일치)는 다음과 같이 정의된다:

$$d(\beta_j) = \mathbb{E}\{-2\log f(y|\beta_j)\}$$

따라서 다음의 관계식이 성립한다:

$$\begin{aligned} 2I(\beta_j) &= \mathbb{E}\{-2\log f(y|\beta_j)\} + \mathbb{E}\{-2\log g(y)\} \\ &= d(\beta_j) + \mathbb{E}\{-2\log g(y)\} \end{aligned}$$

$g(y)$는 β_j에 의존하지 않기 때문에 $I(\beta_j)$를 대신하여 $d(\beta_j)$를 사용한다. $d(\hat{\beta}_j) = \mathbb{E}\{-2\log f(y|\beta_j)\}_{\beta_j=\hat{\beta}_j}$ 역시 참 모델과 적합된 모델 사이의 차이를 근사적으로 반영할 수 있다. 그러나 $d(\hat{\beta}_j)$를 모델 선택에 직접 사용할 수는 없다. $-2\log f(y|\beta_j)$는 $d(\hat{\beta}_j)$의 편향 추정량 (bias estimator)이며, 편향이 β_j의 차원의 두 배만큼 점근적으로 추정될 수 있다. 따라서 $AIC = -2\log f(y|\beta_j) + 2\beta_j$를 정의한다. AIC는 표본 크기 (sample size) n이 변수의 수보다 큰 상황 ($n > p$)에서 $d(\hat{\beta}_j)$의 점근적 불편 추정량이 된다.

BIC

BIC는 모델의 차원을 결정하기 위한 또 다른 방법으로 AIC의 대안으로 쓰인다. BIC의 경우 패널티 항이 두 모델의 Bayes factor로 만들어진다.

식의 전개를 위해 두 모형 $f(y|\beta_1)$와 $f(y|\beta_0)$을 고려하고 $f(y|\beta_1)$에서 m_1을 $f(y|\beta_0)$에서 m_0을 갖는다고 하자. 그리고 $g(\beta_i)$을 $M_i(i=0,1)$이 조건부일때 β_i의 사전분포 (prior density)라 하자. 그러면 bayes factor는 다음과 같다:

$$B_{01}(y) = \frac{m_0(y)}{m_1(y)}$$

여기서 $m_i = \int f(y|\beta_i)g(\beta_i)d\beta_i$이다. 2차 테일러 전개에 의해 m_i을 최대가능도 추정량 (MLE; maximum likelihood estimator)인 $\hat{\beta}_i$로 근사할 수 있다. H_{β_i}을 관측 데이터에 의한 피셔 정보량 (fisher information)이라고 하면 다음의 식을 얻는다:

$$\log\{f(y|\beta_i)g_i(\beta_i)\} \approx \log\{f(y|\hat{\beta}_i)g_i(\hat{\beta}_i)\} - \frac{1}{2}(\beta - \hat{\beta}_i)^T H_{\beta_i}(\beta - \hat{\beta}_i)$$

bayes factor에 적용하면

$$\begin{aligned} m_i(y) &\approx f(y|\hat{\beta}_i)g_i(\hat{\beta}_i) \int \exp\left(-\frac{1}{2}(\beta - \hat{\beta}_i)^T H_{\beta_i}(\beta - \hat{\beta}_i)\right) d\beta_i \\ &= f(y|\hat{\beta}_i)g_i(\hat{\beta}_i)(2\pi)^{\frac{p_i}{2}} n^{-\frac{p_i}{2}} |H_{\beta_i}^{-1}|^{\frac{1}{2}} \end{aligned}$$

를 얻는다. 여기서 p_i는 파라미터 벡터의 차원이다.

그러면

$$\begin{aligned} 2\ln(B_{01}(y)) &= 2\log\frac{m_0(y)}{m_1(y)} \\ &= 2\log\left(\frac{f(y|\hat{\beta}_0)g_0(\hat{\beta}_0)(2\pi)^{\frac{p_0}{2}} n^{-\frac{p_0}{2}} |H_{\beta_0}^{-1}|^{\frac{1}{2}}}{f(y|\hat{\beta}_1)g_1(\hat{\beta}_1)(2\pi)^{\frac{p_1}{2}} n^{-\frac{p_1}{2}} |H_{\beta_1}^{-1}|^{\frac{1}{2}}}\right) \\ &\approx 2\log\left(\frac{f(y|\hat{\beta}_0)}{f(y|\hat{\beta}_1)} + \log\frac{g_0(\hat{\beta}_0)}{g_1(\hat{\beta}_1)} - (p_0 - p_1)\log\left(\frac{n}{2\pi}\right) + \log\left(\frac{H_{\beta_0}^{-1}}{H_{\beta_1}^{-1}}\right)\right) \end{aligned}$$

의 식이 유도된다. 이는 $2\ln(B_{01}(y)) \approx 2\log\left(\frac{f(y|\hat{\beta}_0)}{f(y|\hat{\beta}_1)}\right) - (p_0 - p_1)\log\left(\frac{n}{2\pi}\right)$로 근사된다. 결과적으로 null model $f(y|\beta_0)$과 적합된 모델 $f(y|\beta_1)$을 비교하는 다음의 공식을 얻는다:

$$\text{BIC} = 2\log f(y|\hat{\beta}) + p_i \log(n)$$

만약 균등 분포를 사전분포로 갖는 (즉, 모든 후보모델이 참 모델일 확률이 같다고 가정) BIC는 $n < p$인 상황에서 너무 많은 변수를 선택하는 경향이 있다.