

The AIC Criterion

Imagine that we have a process which is characterised by the location parameter θ and the dispersion parameter σ^2 , Let $x = [x_1, \dots, x_T]'$ and $y = [y_1, \dots, y_T]'$ be mutually independent vectors of T elements drawn from the process, and let $\sigma^2\Omega$, with $\Omega = \Omega(\theta)$, be the dispersion matrix of these vectors. Then, assuming that the process is normal, the likelihood function associated with x is

$$L(\theta, \sigma^2; x) = (2\pi\sigma^2)^{-T/2} |\Omega|^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} S_x \right\}, \quad (1)$$

where $S_x = S(\theta; x)$. The log-likelihood function for x is

$$\log L(\theta, \sigma^2; x) = -\frac{T}{2} \log(2\pi\sigma^2) - \frac{1}{2} \log |\Omega| - \frac{1}{2\sigma^2} S_x, \quad (2)$$

and the corresponding function for y is

$$\log L(\theta, \sigma^2; y) = -\frac{T}{2} \log(2\pi\sigma^2) - \frac{1}{2} \log |\Omega| - \frac{1}{2\sigma^2} S_y. \quad (3)$$

Therefore

$$-2 \log L(\theta, \sigma^2; y) = -2 \log L(\theta, \sigma^2; x) + \frac{1}{\sigma^2} \{S_y - S_x\}. \quad (4)$$

The aim is to derive an expected value of the likelihood or equivalently of $-2 \log(\text{likelihood})$ which is associated with a sample of size T . If this were done for a variety of models described by different restrictions on the parameters vector θ , then we should have a means of discriminating amongst the models.

As a first approximation to the expected value, one might take the probability measure $N(x; \theta, \sigma^2)$ and evaluate it at the point $(\hat{\theta}_x, \hat{\sigma}_x^2)$, where $\hat{\theta}_x = \hat{\theta}(x)$ and $\hat{\sigma}_x^2 = \hat{\sigma}^2(x)$ are estimates of the parameters based on a sample of size T which is provided by the vector x . The problem is that, by and large, the model which imposes the least number of restrictions on the parameters would be the one selected according to the criterion of maximising the likelihood or minimising $-2 \log(\text{likelihood})$.

It is misleading to judge the worth of a model by the extent to which it explains the sample from which its own parameters have been inferred. Instead, it should be judged according to how well it explains the values of other samples. Thus, if the model is estimated from x , then we should test its ability to explain the values in a generic sample y . That is to say, we should find the expected value of the criterion relative to the distribution of y and, at the same time, we should take account of the statistical nature of the sample x which gives rise to the estimates $\hat{\theta}_x$ and $\hat{\sigma}_x^2$.

The argument of Akaike is as follows. Take an estimate $\hat{\theta}_x$ and the corresponding estimate $\hat{\sigma}_x^2 = \hat{\sigma}(\hat{\theta}_x)$ which are based on the information in x and regard these, for the moment, as representing the true model. (there will be other estimates of θ derived from different restrictions which will also be given the same treatment). Now take expectations in the expression $S_y(\hat{\theta}_x) - S_x(\hat{\theta}_x)$. Given that $E\{S_x(\hat{\theta}_x)\} = E\{S_y(\hat{\theta}_y)\}$, where $\hat{\theta}_y$ is the estimator based on y , it follows that

$$E\{S_y(\hat{\theta}_x) - S_x(\hat{\theta}_x)\} = E\{S_y(\hat{\theta}_x) - S_y(\hat{\theta}_y)\}. \quad (5)$$

By taking a Taylor series expansion, we obtain the approximation

$$S_y(\hat{\theta}_x) \simeq S_y(\hat{\theta}_y) + \frac{1}{2}(\hat{\theta}_x - \hat{\theta}_y)' \left[\frac{\partial^2 S_y(\theta)}{\partial \beta_i \partial \beta_j} \right] (\hat{\theta}_x - \hat{\theta}_y). \quad (6)$$

Moreover, given that x and y are mutually independent, we have

$$\begin{aligned} S_y(\hat{\theta}_x) - S_y(\hat{\theta}_y) &\simeq \frac{1}{2}(\hat{\theta}_x - \hat{\theta}_y)' \left[\frac{\partial^2 S_y(\theta)}{\partial \beta_i \partial \beta_j} \right] (\hat{\theta}_x - \hat{\theta}_y) \\ &\simeq \sigma_0^2 (\hat{\theta}_x - \hat{\theta}_y)' Q(\theta_0) (\hat{\theta}_x - \hat{\theta}_y) \\ &\simeq \sigma_0^2 (\hat{\theta}_x - \theta_0)' Q(\theta_0) (\hat{\theta}_x - \theta_0) \\ &\quad + \sigma_0^2 (\theta_0 - \hat{\theta}_y)' Q(\theta_0) (\theta_0 - \hat{\theta}_y), \end{aligned} \quad (7)$$

where $Q(\theta_0)$ is the information matrix which is the expectation of the matrix of second-order partial derivatives evaluated at the true value θ_0 and where σ_0^2 is also a true value. Then, as an asymptotic approximation, we have

$$\begin{aligned} E\{S_y(\hat{\theta}_x) - S_x(\hat{\theta}_x)\} &\simeq \sigma_0^2 E(\hat{\theta}_x - \theta_0)' Q(\theta_0) (\hat{\theta}_x - \theta_0) \\ &\quad + \sigma_0^2 E(\hat{\theta}_y - \theta_0)' Q(\theta_0) (\hat{\theta}_y - \theta_0) \\ &= 2\sigma_0^2 r, \end{aligned} \quad (8)$$

since the quadratic forms are distributed asymptotically as independent chi-square variates with a number of degrees of freedom equal to the number r of free parameters in the estimates.

It remains to evaluate the term $-2 \log L(\theta, \sigma^2; x)$ which is also found in (3). When this is evaluated at $\theta = \hat{\theta}_x$ and $\sigma^2 = \hat{\sigma}_x^2 = S_x/T$, we have

$$-2 \log L(\hat{\theta}_x, \hat{\sigma}_x^2; x) = T \log(2\pi) + T \log \hat{\sigma}_x^2 + \log |\Omega(\hat{\theta}_x)| + T. \quad (9)$$

Taking this result and the result under (7) back to (4) enables us to write

$$-2 \log L(\theta, \sigma^2; y) \simeq c + T \log(\hat{\sigma}_x^2) + 2r, \quad (10)$$

where the term c , which may be disregarded, includes $\log |\Omega(\hat{\theta}_x)|$. This gives us the information criterion:

$$AIC = T \log(\hat{\sigma}_x^2) + 2r. \quad (11)$$