All About AIC

Marco Del Giudice (2019, v. 5)

marcodg@unm.edu

The *Akaike Information Criterion* (AIC) is a versatile criterion for model comparison and model selection. AIC can be used to compare nested or non-nested models, as long as they have been fitted to the same dataset. The form of AIC is a penalized likelihood, with a "penalty" term that depends on the number of free parameters in the model. The meaning of AIC can be understood from three complementary perspectives—that of information theory, that of predictive accuracy, and that of Bayesian statistics.

Information theory

From an information-theoretic perspective, AIC is a data-based estimate of the *Kullback-Leibler distance* between a model *i* and the "true" model that generated the data. The Kullback-Leibler distance, in turn, is the amount of information lost when model *i* is used to approximate the true model. The "best" model in a set can be defined as the one which loses the minimum amount of information, that is, the one characterized by the minimum Kullback-Leibler distance. AIC is an estimate of this information-loss distance for statistical models: thus, *minimizing* AIC means finding the model that *maximizes* the information extracted from data. However, the estimate of the K-L distance provided by AIC is biased by a constant amount, which is usually unknown since one does not know the true model (in many cases, such a model may not even exist in a literal sense). Thus, by computing AIC one cannot know the *absolute* distance between a model and the truth; but if the AIC values of two or more models are compared, the unknown constant drops out and it is possible to rank and compare the models according to their *relative* K-L distance (see Anderson, 2008; Burnham & Anderson, 2002). In this sense, AIC attempts to select models that are *good approximations* but not necessarily true.

Predictive accuracy

From a different perspective, the goal of AIC is to minimize the expected prediction error of a model, that is, maximize its *predictive accuracy*. Specifically, selecting the model with the lowest AIC approximates leave-one-out cross-validation as sample size approaches infinity. Note that the true model may not be the one that maximizes predictive accuracy; again, AIC selects models that are good approximations (small prediction error) but not necessarily true. For the same reason, AIC privileges predictive accuracy over parsimony, in contrast with the *Bayesian Information Criterion* (BIC, another penalized likelihood criterion), which attempts to select the true model if such model is present in the comparison set. AIC tends to select more complex models than BIC, particularly when sample size is large (see Aho et al., 2014; Weakliem, 2016).

Bayesian interpretation

From a Bayesian perspective, AIC approximates a Bayes factor (BF) in which the prior changes with sample size. Specifically, the prior implied by AIC is based on a

fixed proportion of the total observations. In contrast, BIC approximates a Bayes factor in which the prior is based on the information from a single observation (*unit information prior*), regardless of sample size. AIC values can be used to estimate Bayes factors and posterior probabilities for the models in a comparison set (Weakliem, 2016; see below).

Basic AIC formulas

AIC is computed from the maximized likelihood of a fitted model ($\mathcal{L}$) and the number of free parameters in the model (K):

$$AIC = -2\ln(\mathcal{L}) + 2K$$

If all the models in the set have normally distributed errors with constant variance, their AIC values can be computed from least-squares statistics (i.e., the standard output of many statistical applications):

$$AIC = N\ln\left(\frac{RSS}{N}\right) + 2K$$

where RSS is the residual sum of squares, N is the sample size, and K is the number of free parameters, including the intercept and σ^2.

When sample size is small with respect to the maximum K in the set ($N/K < 40$ is a reasonable rule of thumb), AIC is no longer adequate, and a better estimate is provided by the second-order AIC or AICc. For linear models with normally distributed residuals, AICc is:

$$AIC_c = -2\ln(\mathcal{L}) + 2K\frac{N}{N-K-1}$$

Equivalently, for the least-squares case:

$$AIC_c = N\ln\left(\frac{RSS}{N}\right) + 2K\frac{N}{N-K-1}$$

The Δ_i statistic of model i is simply

$$\Delta_i = AIC_i - AIC_{min}$$

where AIC_{min} is the minimum value of AIC in the set (hence, $\Delta_i = 0$ for the best model). These are common rules of thumb for model comparison (see Burnham & Anderson, 2002):

- models with Δ_i less than 2 perform about as well as the best model in the set;
- models with Δ_i between 2 and 4 receive considerable support from the data;
- models with Δ_i between about 4 and 7 have non-negligible support;
- models with Δ_i larger than about 10 have essentially no support.

Note: a model with $\Delta_i \approx 2$ that only differs from the best model by one additional parameter is *not* supported by the data: since adding one parameter increases AIC by

exactly 2, the model has failed to improve the fit to the data (same maximized likelihood), and has been penalized for having one parameter in excess.

Model probabilities (Akaike weights)

The *Akaike weight* of model *i* is computed as follows:

$$w_i = \frac{e^{-\Delta_i/2}}{\sum e^{-\Delta_i/2}}$$

where Σ is the summation over all the models in the set. Note that values of Δ_i depend on the best model in the set, and values of w_i are relative to a particular set of models. A model's Akaike weight can be interpreted as the probability of that model *actually* being the best model in the set. Thus, Akaike weights directly quantify the uncertainty associated with model selection, in contrast with the simple ranking of models from best to worst. The relative amount of evidence favoring one model over another can be easily quantified by computing an *evidence ratio*, i.e., the ratio of the two model probabilities (Burnham & Anderson, 2002).

From a Bayesian perspective, $e^{-\Delta_i/2}$ approximates a Bayes factor that compares model *i* with the best model in the set (note that the priors implied by AIC are different from those implied by BIC; see above). Thus, $\Delta_i = 2$ corresponds to $BF_i \approx 0.37$ (model *i* is about 37% as likely as the best model in the set); $\Delta_i = 4$ corresponds to $BF_i \approx 0.14$; and $\Delta_i = 10$ corresponds to $BF_i \approx 0.007$. The Akaike weight of a model is simply its posterior probability, assuming equal priors for all the models in the set. See Weakliem (2016) for more details.

When the same variables recur in multiple models, a simple index of the relative importance of a variable *j* can be obtained by summing the w_i of all models in the set that contain that variable. The "cumulative" weight of variable *j* is indicated by $w_+(j)$. Variables with higher w_+ are more strongly supported by the evidence than variables with lower values of w_+. In this way, AIC can be used to rank individual variables in addition to models.

References

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology, 95,* 631-636.

Anderson, D. R. (2008). *Model based inference in the life sciences: A primer on evidence*. New York: Springer.

Burnham, K.P. & Anderson, D.R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach.* Springer.

Weakliem, D. L. (2016). *Hypothesis testing and model selection in the social sciences*. Guilford.