Akaike Information Criterion

The Akaike information criterion (AIC*)* was first developed by Akaike (1973) to compare different models on a given outcome. For example, if researchers are interested in what influences the rating of a wine, they can make all the influencing factors and make them independent variables. Then see how these variables influence the rating of a wine, one may estimate several different regression models using one or more independent variables.

These models are more than just a tool that can reveal the relationship between the outcomes and a specific variable(s). Model selection is important for several reasons. If there is under-fitting the model could be missing of fail to accurately identify the true nature of variability in the outcome variable(s). On the other hand, an over-fitted model tends to lose generality. AIC and later to be discused BIC Are ways to select a model that best balances these drawbacks.

Understand these methods will not tell you which model in the universe will best represent the data. They can only identify which model out of a presented set of models. If the presented models are all bad it will identify the least bad, or best one in the bunch. For this reason, it is strongly sugested that once a best model is determined, the modle is verified using traditional null-hypothesis testing.

Akaike (1973) showed that the "best" model is determined by an AIC score:

$$\mathbf{AIC = 2K - 2log\ (\mathcal{L}\ \ (\theta \mid y)),}$$

Where:

- *K*= the number of estimable parameters (degrees of freedom)
- Log ($\mathcal{L}$ (θ | y)) = the log-likelihood at its maximum point of the model estimated.

(Burnham and Anderson, (2002), Hurvich and Tsai (1989)) further refined the AIC estimate to correct for small data samples:

$$\mathbf{AICc = AIC + 2K\ (K + 1)\ /\ n - K - 1,}$$

Where:

- *N* = the sample size.
- *K*= the number of estimable parameters (degrees of freedom).
- AIC is defined above.

If *n* is large with respect to *K*, then the original AIC formula is sufficient. AICc is more general, used in place of AIC when *n* is small with respect to *K*. The best model is then the model with the lowest *AICc* (or AIC) score.

NOTE: that the *AIC* and *AICc* scores are ordinal and mean nothing on their own. They are simply a way of ranking the a given set of models.

Burnham & Anderson (2001) suggest using AIC and other information-theoretical approaches to settle conflicts in the applied sciences. Still ther are advantages and disadvantages to using AIC.

AIC Advantages

1. AIC can objectively determins which model among a set of models is better than the rest.
2. It is easy for software calculate and interpret.
3. The AIC supplies information on the strength of evidence for each model. making
4. concept of significance becomes superfluous with the AIC.
5. In the case where there are many models ranked highly based on the AIC, model uncertainty can be incorporate to obtain robust estimates, and confidence intervals.
6. It penalizes multiple paramiters to guard against over-filling.

AIC Limitations:

- a model is only good as the data which have generated it.
- conclusions depend on the set of candidate models specified before the analyses is conducted

Bayesian Information Criterion

The Bayesian information criterion (BIC) or Schwarz criterion (SBC, SBIC) is a criterion for model selection developed by Schwarz (1978) Schwarz presented a Bayesian argument for adopting the use of the model. In the 2012 paper "'All models are wrong...': an introduction to model uncertainty" Wit, Heuvel, & Romeijn, J. W. (2012).). Discribe the BIC formula as:

$$\boldsymbol{BIC} = \mathbf{ln}(\boldsymbol{n})\,\boldsymbol{k} - \mathbf{2ln}\ (\acute{\mathrm{L}})$$

Where:

- $\acute{\mathrm{L}}$ = the maximized value of the likelihood function of the model M, i.e.
- $\acute{\mathrm{L}}\ = p(x|\theta, M)$, where θ are the parameter values that maximize the likelihood function;
- x = the observed data;
- n = the number of data points in x, the number of observations, or equivalently, the sample size;
- k = the number of parameters estimated by the model. If the model under consideration is a multiple linear regression, then the estimated parameters are the intercept, q slope parameters and the constant variance of the errors. Thus, $k = q + 2$;

The BIC is an asymptotic result derived under the assumptions that the data distribution is in an exponential family. That is, the integral of the likelihood function $p(x|\theta, M)$ times the prior probability distribution $p(\theta|M)$, over the parameters θ of the model M for fixed observed data x is approximated as:

$$-\mathbf{2} * \mathbf{ln}\,\boldsymbol{p}(\boldsymbol{x}|\boldsymbol{M}) \approx \boldsymbol{BIC} = \ -\mathbf{2} * \mathbf{ln}\,\acute{\mathrm{L}} + \boldsymbol{k} * \left(\mathbf{ln}(\boldsymbol{n})_{\mathbf{ln}(\mathbf{2}\Pi)}\right).$$

For large n , this can be approximated by the formula given above.

There are advantages and limitations for using BIC.

BIC advantages:

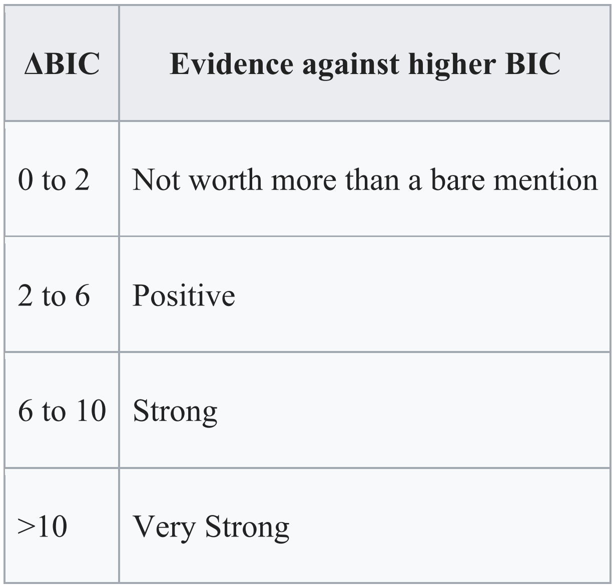
1. Is independent of the prior
2. Can measure the efficiency of the parameterized model in terms of predicting the data.
3. Penalizes the complexity of the model where complexity refers to the number of parameters in the model.
4. Is approximately equal to the minimum description length criterion but with negative sign.

5. Can be used to choose the number of clusters according to the intrinsic complexity present in a given dataset.
6. Is closely related to other penalized likelihood criteria like AIC
7. It penalizes multiple paramiters to guard against under-filling.

BIC limitations:

1. the above approximation is only valid for sample size n much larger than the number k of parameters in the model.
2. the BIC cannot handle complex collections of models as in the variable selection problem in high-dimension.

The strength of the evidence against the model with the higher BIC value can be summarized as follows



ΔBIC	Evidence against higher BIC
0 to 2	Not worth more than a bare mention
2 to 6	Positive
6 to 10	Strong
>10	Very Strong

References

Akaike H. (1973) *Information theory as an extension of the maximum likelihood principle B.N. Petrov, F. Csaki* (Eds.), Second International Symposium on Information Theory, Akademiai Kiado, Budapest (1973), pp. 267-281

Burnham, K., Anderson, D (2002) *Model Selection and Multimodal Inference.*

Springer, New York

Burnham, K. P., & Anderson, D. R. (2001). Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife research, 28*: 111-119.

Hurvich, C., Tsai, C.-L. (1989) Regression and time series model selection in small samples. *Biometrica, 76:* 297–293

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics, 6:* 461-464.

Wit, E., Heuvel, E. V. D., & Romeijn, J. W. (2012). ‘All models are wrong...’: an introduction to model uncertainty. *Statistica Neerlandica, 66*: 217-236.