The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications.
Frank J. Fabozzi, Sergio M. Focardi, Svetlozar T. Rachev and Bala G. Arshanapalli.

APPENDIX E

Model Selection Criterion: AIC and BIC

In several chapters we have discussed goodness-of-fit tests to assess the performance of a model with respect to how well it explains the data. However, suppose we want to select from among several candidate models. What criterion can be used to select the best model? In choosing a criterion for model selection, one accepts the fact that models only approximate reality. Given a set of data, the objective is to determine which of the candidate models best approximates the data. This involves trying to minimize the loss of information. Because the field of information theory is used to quantify or measure the expected value of information, the information-theoretic approach is used to derive the two most commonly used criteria in model selection—the Akaike information criterion and the Bayesian information criterion.[1] These two criteria, as described in this appendix, can be used for the selection of econometric models.[2]

[1]There are other approaches that have been developed. One approach is based on the theory of learning, the Vapnik-Chervonenkis (VC) theory of learning. This approach offers a complex theoretical framework for learning that, when applicable, is able to give precise theoretical bounds to the learning abilities of models. Though its theoretical foundation is solid, the practical applicability of the VC theory is complex. It has not yet found a broad following in financial econometrics. See Vladimir N. Vapnik, *Statistical Learning Theory* (New York: John Wiley & Sons, 1998).

[2]For a further discussion of these applications of AIC and BIC, see Herman J. Bierens, "Information Criteria and Model Selection," Pennsylvania State University, March 12, 2006, working paper. In addition to the AIC and BIC, Bierens discusses another criterion, the Hannan-Quinn criterion, in E. J. Hannan and B. G. Quinn, "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society* B, no. 41 (1979): 190–195.

AKAIKE INFORMATION CRITERION

In 1951, Kullback and Leibler developed a measure to capture the information that is lost when approximating reality; that is, the Kullback and Leibler measure is a criterion for a good model that minimizes the loss of information.[3] Two decades later, Akaike established a relationship between the Kullback-Leibler measure and maximum likelihood estimation method—an estimation method used in many statistical analyses as described in Chapter 13—to derive a criterion (i.e., formula) for model selection.[4] This criterion, referred to as the *Akaike information criterion* (AIC), is generally considered the first model selection criterion that should be used in practice. The AIC is

$$\text{AIC} = -2\log L(\hat{\theta}) + 2k$$

where θ = the set (vector) of model parameters
$L(\hat{\theta})$ = the likelihood of the candidate model given the data when evaluated at the maximum likelihood estimate of θ
k = the number of estimated parameters in the candidate model

The AIC in isolation is meaningless. Rather, this value is calculated for every candidate model and the "best" model is the candidate model with the smallest AIC. Let's look at the two components of the AIC. The first component, $-2\log L(\hat{\theta})$, is the value of the likelihood function, $\log L(\hat{\theta})$, which is the probability of obtaining the data given the candidate model. Since the likelihood function's value is multiplied by –2, ignoring the second component, the model with the minimum AIC is the one with the highest value for the likelihood function. However, to this first component we add an adjustment based on the number of estimated parameters. The more parameters, the greater the amount added to the first component, increasing the value for the AIC and penalizing the model. Hence, there is a trade-off: the better fit, created by making a model more complex by requiring more parameters, must be considered in light of the penalty imposed by adding more parameters. This is why the second component of the AIC is thought of in terms of a penalty.

---

[3] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *Annals of Mathematical Statistics* 22, no. 1 (1951): 79–86.

[4] Hirotugu Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, ed. B. N. Petrov and F. Csake (Budapest: Akademiai Kiado, 1973), 267–281; and Hirotugu Akaike, "A New Look at the Statistical Model Identification," *I.E.E.E. Transactions on Automatic Control, AC* 19, (1974): 716–723.

For small sample sizes, the *second-order Akaike information criterion* (AIC_c) should be used in lieu of the AIC described earlier. The AIC_c is

$$AIC_c = -2\log L(\hat{\theta}) + 2k + (2k+1)/(n-k-1)$$

where n is the number of observations.[5] A small sample size is when n/k is less than 40. Notice as the n increases, the third term in AIC_c approaches zero and will therefore give the same result as AIC. AIC_c has also been suggested to be used instead of AIC when n is small or k is large.[6] It has been suggested, for example, that in selecting the orders of an ARMA, as we described in Chapter 9, the AIC_c be used.[7]

Typically, to assess the strength of evidence for the each candidate model, two measures can be used:

1. The delta AIC
2. The Akaike weights

Consider first the *delta AIC* measure assuming there are M candidate models. An AIC can be calculated for each candidate model, denoted by AIC_m ($m = 1, \ldots, M$). The AIC with the minimum value, denoted by AIC^*, is then the best model. The delta AIC for the mth candidate model, denoted by Δ_m, is simply the difference between the AIC_m and AIC^*. This difference is then used as follows to determine the level of support for each candidate model. If the delta AIC is

- Less than 2, this indicates there is substantial evidence to support the candidate model (i.e., the candidate model is almost as good as the best model).
- Between 4 and 7, this indicates that the candidate model has considerably less support.
- Greater than 10, this indicates that there is essentially no support for the candidate model (i.e., it is unlikely to be the best model).[8]

The above values for the computed delta AICs are merely general rules of thumb.

Because the magnitude of the delta AIC is not meaningful in itself, to measure the strength of evidence for a candidate model we are interested

[5] Clifford M. Hurvich and Chih-Ling Tsai, "Regression and Time Series Model Selection in Small Samples," *Biometrika* 76, no. 2 (June 1989): 297–307.

[6] Kenneth P. Burnham and David R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach,* 2nd ed. (New York: Springer-Verlag, 2002).

[7] Peter J. Brockwell and Richard A. Davis, *Time Series: Theory and Methods*, 2nd ed. (New York: Springer-Verlag, 2009), 273.

[8] Burnham and Anderson, *Model Selection and Multimodel Inference*, 70.

in the relative value of the delta AIC. The *Akaike weights*, denoted by w_m, are obtained by normalizing the relative likelihood values. That is, they are the ratios of a candidate model's delta AIC relative to the sum of the delta AICs for all candidate models, shown as follows:

$$w_m = \frac{\exp(-0.5\Delta_m)}{\sum_{j=1}^{M} \exp(-0.5\Delta_j)}$$

The interpretation of this measure of strength of each candidate model given the data is the following: the Akaike weights are the probability that the candidate model is the best among the set of candidate models. For example, if a candidate model has an Akaike weight of 0.60, this means that given the data, the candidate model has a 60% probability of being the best one.

Further information can be obtained by calculating the ratio of Akaike weights for different candidate models to determine to what extent one candidate model is better than another candidate model. These measures, called *evidence ratios*, can be used to compare, for example, the best model versus a candidate model. For example, if the evidence ratio computed as the ratio of the best model to some candidate model is 1.8, then this can be interpreted as the best model being 1.8 times more likely than that candidate model of being the best model.

What is the difference between the AIC and hypothesis tests in model selection described in Chapters 3 and 4 where we described statistical tests for various regression models and the use of stepwise regressions? The difference is that in those earlier chapters, the tests used for model selection are hypothesis tests where at a certain level of confidence an independent variable would be included or excluded from the model. In contrast, model selection applying AIC is based on the strength of the evidence and provides for each of the candidate models a measure of uncertainty. What is important to emphasize is that the AIC might identify which model is best among the candidate models but that does not mean that any of the candidate models do a good job of explaining the data.

BAYESIAN INFORMATION CRITERION

The *Bayesian information criterion* (BIC), proposed by Schwarz[9] and hence also referred to as the *Schwarz information criterion* and *Schwarz Bayesian*

[9]Gideon Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics* 6 (1978): 461–464. The purpose of the BIC is to provide an asymptotic approximation to a transformation of the candidate model's Bayesian posterior probability.

information criterion, is another model selection criterion based on information theory but set within a Bayesian context. The difference between the BIC and the AIC is the greater penalty imposed for the number of parameters by the former than the latter. Burnham and Anderson provide theoretical arguments in favor of the AIC, particularly the AIC_c over the BIC.[10] Moreover, in the case of multivariate regression analysis, Yang explains why AIC is better than BIC in model selection.[11]

The BIC is computed as follows:

$$\text{BIC} = -2\log L(\hat{\theta}) + k\log n$$

where the terms above are the same as described in our description of the AIC.

The best model is the one that provides the minimum BIC, denoted by BIC*. Like delta AIC for each candidate model, we can compute delta $\text{BIC} = \text{BIC}_m - \text{BIC}^*$. Given M models, the magnitude of the delta BIC can be interpreted as evidence *against a candidate model* being the best model. The rules of thumb are[12]

- Less than 2, it is not worth more than a bare mention.
- Between 2 and 6, the evidence against the candidate model is positive.
- Between 6 and 10, the evidence against the candidate model is strong.
- Greater than 10, the evidence is very strong.

[10] Burnham and Anderson, *Model Selection and Multimodel Inference*.

[11] Ying Yang, "Can the Strengths of AIC and BIC Be Shared?" *Biometrika* 92, no. 4 (December 2005): 937–950.

[12] Robert E. Kass and Adrian E. Raftery, "Bayes Factors," *Journal of the American Statistical Association* 90, no. 430 (June 1995): 773–795. The rules of thumb provided here are those modified in a presentation by Joseph E. Cavanaugh, "171:290 Model Selection: Lecture VI: The Bayesian Information Criterion" (PowerPoint presentation, The University of Iowa, September 29, 2009).