[2] Hu, X. and Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, **59**, 21–47.

[3] Klauer, K. C. (1996). Urteilerübereinstimmung für dichotome Kategoriensysteme. *Diagnostica*, **42**, 101–118.

[4] Klauer, K. C. and Batchelder, W. H. (1996). Structural analysis of subjective categorical data. *Psychometrika*, **61**, 199–240.

(CATEGORICAL DATA, SUBJECTIVE
KAPPA COEFFICIENT
MEASURES OF AGREEMENT)

KARL CHRISTOPH KLAUER

AIC *See* AKAIKE'S INFORMATION CRITERION

AKAIKE'S INFORMATION CRITERION II

AIC (*an information criterion*, or *Akaike's information criterion*) is a statistic defined for parametric models whose parameters have been obtained by maximizing a form of likelihood* function. AIC values are compared in selecting from among competing models for a data set used for parameter estimation. The selection is prescribed by Akaike's minimum AIC criterion, hereafter MinAIC, which says that the model with smallest AIC is to be preferred [1, 2, 3].

Consider a model family with real parameter vector $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_p)$ specifying a candidate family of joint probability density functions $L_N(\boldsymbol{\theta}; x_1, \ldots, x_N)$, $\boldsymbol{\theta} \in \Theta$, for observations $x_1, \ldots, x_N$ of the random variables $X_1, \ldots, X_N$. Suppose $L_N(\boldsymbol{\theta}) = L_N(\boldsymbol{\theta}; x_1, \ldots, x_N)$ is maximized over Θ at $\hat{\boldsymbol{\theta}}_N = \hat{\boldsymbol{\theta}}_N(x_1, \ldots, x_N)$ satisfying

$$\frac{\partial}{\partial \boldsymbol{\theta}} L_N(\theta) \Big|_{\theta = \hat{\theta}_N} = 0 \quad (1)$$

(*see* MAXIMUM LIKELIHOOD ESTIMATION). Then the AIC of the model for $X_1, \ldots, X_N$ determined by $\hat{\boldsymbol{\theta}}_N$ is

$$\mathrm{AIC}_N(\hat{\boldsymbol{\theta}}_N) = -2 \ln L_N(\hat{\boldsymbol{\theta}}_N) + 2 \dim \boldsymbol{\theta}\,, \quad (2)$$

where $\dim \boldsymbol{\theta} = p + 1$, $p \geqslant 0$. The minimum-AIC choice can be determined from the signs of the differences of AIC values. Therefore, only properties of differences of AIC values are important, not the AIC values themselves. In particular, for comparing any two competing model families $L_N^{(i)}(\boldsymbol{\theta}^{(i)}; x_1, \ldots, x_N)$, $\boldsymbol{\theta}^{(i)} \in \Theta^{(i)}$, $i = 1, 2$, with parameter estimates $\hat{\boldsymbol{\theta}}_N^{(1)}$ and $\hat{\boldsymbol{\theta}}_N^{(2)}$, respectively, the properties of the minimum AIC criterion, and their practical consequences, can be determined from properties of

$$\mathrm{AIC}_N(\hat{\theta}_N^{(1)}) - \mathrm{AIC}_N(\hat{\theta}_N^{(2)}) = -2 \ln \frac{L_N^{(1)}(\hat{\theta}_N^{(1)})}{L_N^{(2)}(\hat{\theta}_N^{(2)})} + 2(\dim \theta^{(1)} - \dim \theta^{(2)})\,. \quad (3)$$

EXTENSIONS OF THE CONCEPT OF LIKELIHOOD FUNCTION FOR AIC

Each family $L_N(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, will be referred to as a likelihood function, but it is important to understand the quite general sense in which this term is used with AIC in order to appreciate the scope of MinAIC. First, the $L_N(\boldsymbol{\theta})$ can be probability density functions in the most general sense. For example, when $X_1, \ldots, X_N$ are discrete-valued, as in the case of categorical data*, they will be the probability functions assigning probabilities to all possible values of $(x_1, \ldots, x_N)$ [18, 17]. [In the language of measure theory*, the $L_N(\boldsymbol{\theta})$ must be probability density functions for some measure, not necessarily Lebesgue measure, with respect to which the probability measure of $X_1, \ldots, X_N$ has a probability density.] Further, the parametric family $L_N(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, is not subject to the traditional requirement that there be a $\boldsymbol{\theta}_0 \in \Theta$ such that $L_N(\boldsymbol{\theta}_0)$ coincides with the true probability density function $g_N(x_1, \ldots, x_N)$ of $X_1, \ldots, X_N$. However, the model family should provide close approximations to the relevant characteristics of $X_1, \ldots, X_N$ in order for the parameter dimension terms on the right in (2) and (3) to play the role desired by Akaike for the large-sample means of AIC differences discussed in the next section.

For example, with regression models* and time-series* models, it is common to use parameter estimates that maximize Gaussian likelihood functions, even when the data are not Gaussian, in order to estimate just their means, variances, and covariances. If $L_N^{(i)}(\boldsymbol{\theta}^{(i)}; x_1, \ldots, x_N)$, $\boldsymbol{\theta}^{(i)} \in \Theta^{(i)}$, $i = 1, 2$, are of Gaussian form and can correctly describe the first and second moments of the data, and if the model (1) is a special case of the model (2), so that dim $\theta^{(1)} <$ dim $\theta^{(2)}$, it happens under rather general non-Gaussian assumptions that the likelihood ratio term in (3) will have the same limiting distribution as $N \to \infty$ that it has with Gaussian data, usually the chi-square distribution* with d.f. = dim $\theta^{(2)}$ − dim $\theta^{(1)}$:

$$-2 \ln \frac{L_N^{(1)}(\hat{\boldsymbol{\theta}}_N^{(1)})}{L_N^{(2)}(\hat{\boldsymbol{\theta}}_N^{(2)})} \simeq \chi^2_{\dim \theta^{(2)} - \dim \theta^{(1)}}. \quad (4)$$

This conclusion can be obtained from Theorem 3 and Lemma 3 of ref. [14] in the case of linear (stationary or suitably orthogonalizable) regression models, and from ref. [5] for some other time series models; see also ref. [16]. Since the chi-square distribution in (4) has mean dim $\boldsymbol{\theta}^{(2)}$ − dim $\boldsymbol{\theta}^{(1)}$, this result and (3) suggest that the means of the AIC differences satisfy

$$\lim_{N\to\infty} E_{X_1,\ldots,X_N}[\mathrm{AIC}_N(\hat{\theta}_N^{(1)}) - \mathrm{AIC}_N(\hat{\theta}_N^{(2)})] = \dim \theta^{(1)} - \dim \theta^{(2)}. \quad (5)$$

So on average MinAIC will select the lower-dimensional and therefore less overparametrized model.

AIC can often be derived for conditional likelihoods when the conditioning variables are the same for all models being compared. This is attractive when the conditional likelihoods are easier to maximize. Consider the case of selecting the order p of an autoregressive model

$$X_t = \theta_1 X_{t-1} + \cdots + \theta_p X_{t-p} + \varepsilon_t \quad (6)$$

for time-series variates $X_1, \ldots, X_{N+P}$ from a range of orders $1 \leqslant p \leqslant P$. For each model, it is assumed that the ε_t have mean zero and constant variance, and are independent of all X_s, $s < t$. Because of the last property, conditioning on $X_1, \ldots, X_P$ produces, for a given p, the conditional Gaussian likelihoods

$$L_N^{(p)}(\boldsymbol{\theta}) = \frac{1}{(2\pi\theta_0)^{N/2}} \exp\left(-\frac{1}{2\theta_0} \times \sum_{t=P+1}^{P+N} (x_t - \theta_1 x_{t-1} - \cdots - \theta_p x_{t-p})^2\right). \quad (7)$$

The maximizing coefficients $\hat{\theta}_j^{(p)}$, $1 \leqslant j \leqslant p$, are the ordinary least squares coefficient estimates minimizing $\sum_{t=P+1}^{P+N}(x_t - \theta_1 x_{t-1} - \cdots - \theta_p x_{t-p})^2$, and subsequent maximization with respect to θ_0 yields

$$\mathrm{AIC}_N(\hat{\boldsymbol{\theta}}_N^{(p)}) = N \ln(2\pi e \hat{\sigma}^2_{N,p}) + 2(p + 1), \quad (8)$$

with $\hat{\sigma}^2_{N,p} = N^{-1} \sum_{t=P+1}^{P+N}(x_t - \hat{\theta}_1^{(p)} x_{t-1} - \cdots - \hat{\theta}_p^{(p)} x_{t-p})^2$. The unconditional Gaussian likelihoods for autoregressive models have a more complex form than $L_N^{(p)}(\boldsymbol{\theta})$ in (7) and require nonlinear methods for the solution of (1) [10]. (For unconditional likelihood functions for time series models, free software is available via the Internet for calculating AICs and a diagnostic for the stability of the MinAIC choice over time [7].)

THEORETICAL PROPERTIES

AICs of the form (8) will be considered first because they occur widely in the regression literature. For N large enough relative to P, the value $\hat{p}_{\mathrm{MinAIC}}$ of p minimizing AIC will coincide with the p minimizing Akaike's final prediction error criterion (sometimes called Akaike's criterion*),

$$\mathrm{FPE}_{N,p} = N\left(\frac{N + p + 1}{N - p - 1}\right)\hat{\sigma}^2_{N,p}. \quad (9)$$

Many properties of this criterion and of (8), also for the case of nonrandom regressors, with $\hat{\sigma}^2_{N,p}$ replaced by the m.l. estimate of regression error variance, are discussed in the ESS, in the entries REGRESSION VARIABLES, SELECTION OF (vol. 7, pp. 709–714); LINEAR MODEL SELECTION, CRITERIA AND TESTS (Supp. pp. 83–87), and GENERALIZED FINAL PREDICTION ERROR CRITERIA (Update vol. 1, pp. 269–272). We do not repeat details here, except to summarize by referring to two properties easily stated for AIC. When the time series X_t being modeled as a finite-order autoregression (6) is, instead, an infinite-order autoregression, (8) has for one-step-ahead prediction an optimality property discovered by Shibata [20] that is not shared by other criteria of the form

$$N \ln 2\pi e \hat{\sigma}^2_{N,p} + C(p + 1), \quad (10)$$

with $C \neq 2$, in particular not by the Schwarz criterion* with $C = \log N$. This property requires P to approach ∞ with N in such a way that $P^2/N \to 0$. On the other hand, if X_t is an autoregressive process of finite order $p_0 < P$, then $\hat{p}_{\mathrm{MinAIC}}$ is an "overconsistent" estimator of p_0 in the sense that $\Pr\{\hat{p}_{\mathrm{MinAIC}} \geqslant p_0\} \to 1$ as $N \to \infty$, but is not consistent [19] except in the infinite variance case [4]. By contrast, the minimizer of (10) consistently estimates p_0 whenever $C \to \infty$ as $N \to \infty$ with $C/N \to 0$.

The conceptual leap from the final prediction error criterion for autoregressions to AIC for general statistical models (2) was made by Akaike in 1971 in the context of comparing factor analysis* models. It is not immediately obvious how to view this as a prediction problem. Akaike's insight, recalled in ref. [8], had two components. First, one can view the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_N(x_1, \ldots, x_N)$ obtained from any parametric family $L_N(\boldsymbol{\theta}; x_1, \ldots, x_N)$, $\boldsymbol{\theta} \in \Theta$, as providing a "prediction" $\overset{*}{L}_N(\hat{\boldsymbol{\theta}}_N) = L_N(\hat{\boldsymbol{\theta}}_N; \overset{*}{x}_1, \ldots, \overset{*}{x}_N)$ of a probability density function for observations $\overset{*}{x}_1, \ldots, \overset{*}{x}_N$ from an independent replicate $\overset{*}{X}_1, \ldots, \overset{*}{X}_N$ of $X_1, \ldots, X_N$ obtained in the future. Second, the goodness of this prediction can be measured by the Kullback information* discrepancy from the true density $g_N(\overset{*}{x}_1, \ldots, \overset{*}{x}_N)$ to $\overset{*}{L}_N(\hat{\boldsymbol{\theta}}_N)$,

$$I(g_N; \overset{*}{L}_N(\hat{\boldsymbol{\theta}}_N)) = E_{X_1^*, \ldots, X_N^*}[\ln g_N] - E_{X_1^*, \ldots, X_N^*}[\ln \overset{*}{L}_N(\hat{\boldsymbol{\theta}}_N)],$$

more specifically by the average discrepancy $E_{X_1, \ldots, X_N}[I(g_N; \overset{*}{L}_N(\hat{\boldsymbol{\theta}}_N))]$. Using the notation $a_N \approx b_N$ to mean $a_N - b_N \to 0$ as $N \to \infty$, the property desired of AIC for any two model families being compared is

$$E_{X_1, \ldots, X_N}[\mathrm{AIC}_N(\hat{\theta}_N^{(1)}) - \mathrm{AIC}_N(\hat{\theta}_N^{(2)})] \approx 2E_{X_1, \ldots, X_N}[I(g_N; L_N^{(1)*}(\hat{\theta}_N^{(1)})) - I(g_N; L_N^{(2)*}(\hat{\theta}_N^{(2)}))]. \quad (11)$$

Then the model with smaller AIC will, on average, be the one whose predicted density has smaller average discrepancy from the true density. Under some regularity conditions, this property is achieved by the definition (2) when each parametric family $L_N(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, has a density $L_N(\boldsymbol{\theta}_0)$ that coincides with g_N (or, in some cases, reproduces the features of g_N being modeled, such as its first and second moments).

To indicate how this comes about, we observe that since

$$I(g_N; L_N^{(1)*}(\hat{\boldsymbol{\theta}}_N^{(1)})) - I(g_N; L_N^{(2)*}(\hat{\boldsymbol{\theta}}_N^{(2)})) = E_{X_1^*, \ldots, X_N^*}[\ln L_N^{(2)*}(\hat{\boldsymbol{\theta}}_N^{(2)}) - \ln L_N^{(1)*}(\hat{\boldsymbol{\theta}}_N^{(1)})],$$

it is enough to verify

$$2E_{X_1, \ldots, X_N}[L_N(\hat{\boldsymbol{\theta}}_N) - E_{X_1^*, \ldots, X_N^*}[\overset{*}{L}_N(\hat{\boldsymbol{\theta}}_N)]] \to 2 \dim \boldsymbol{\theta} + K \quad (12)$$

for some constant K that is the same for all models being compared. Because $E_{X_1, \ldots, X_N}[\ln L_N(\boldsymbol{\theta}_0)] = E_{X_1^*, \ldots, X_N^*}[\ln \overset{*}{L}_N(\boldsymbol{\theta}_0)]$, the left-hand side of (12) has the decomposition

$$\begin{aligned} 2E_{X_1, \ldots, X_N}[\ln L_N(\hat{\boldsymbol{\theta}}_N) - E_{X_1^*, \ldots, X_N^*}[\ln \overset{*}{L}_N(\hat{\boldsymbol{\theta}}_N)]] &= 2E_{X_1, \ldots, X_N}[\ln L_N(\hat{\boldsymbol{\theta}}_N) - \ln L_N(\boldsymbol{\theta}_0)] \\ &\quad + 2E_{X_1, \ldots, X_N}[E_{X_1^*, \ldots, X_N^*}[\ln \overset{*}{L}_N(\boldsymbol{\theta}_0)] \\ &\quad - E_{X_1^*, \ldots, X_N^*}[\ln \overset{*}{L}_N(\hat{\boldsymbol{\theta}}_N)]], \end{aligned} \quad (13)$$

and it suffices to show that each of the two terms on the right tends to

$$\dim \theta + K/2. \quad (14)$$

As $\boldsymbol{\theta}_0$ is the maximizer of $E_{X_1^*, \ldots, X_N^*}[\ln \overset{*}{L}_N(\boldsymbol{\theta})]$ [the minimizer of $I(g_N; \overset{*}{L}_N(\boldsymbol{\theta}))$], one will usually have

$$\frac{\partial}{\partial \boldsymbol{\theta}} E_{X_1^*, \ldots, X_N^*}[\ln \overset{*}{L}_N(\boldsymbol{\theta})] \bigg|_{\boldsymbol{\theta} = \boldsymbol{\theta}_0} = 0. \quad (15)$$

It follows from this and from (1) that, in the second-order Taylor expansions* of the terms inside the expectations on the right in (13) about $\hat{\boldsymbol{\theta}}_N$ and $\boldsymbol{\theta}_0$ respectively, only the second-order terms are nonzero. The analysis of these and their means leads to (14) for each expansion [1, 18, 5, 21]. In the case of (8) for a stationary autoregressive process of order p_0 whose error process ε_t has variance σ^2 and fourth cumulant κ_4, the constant K in (12) has the value κ_4/σ^4 [5].

GENERALIZATIONS

A variety of generalizations of AIC have been proposed in which $\dim \boldsymbol{\theta}$ in (2) is replaced by an estimate of the left-hand side of (12) for small N [22, 11, 9]; or it is replaced by the limit of this quantity when (12) fails because of modifications to the likelihood or because the model family is incorrect [21, 5, 15]. The last reference also considers analogues of AIC when functions other than likelihoods are optimized to estimate parameters.

Recent research has focused on generalizations to obtain (11) when the parameter estimates at which the log likelihoods are evaluated are not maximum likelihood estimates but, say, robust estimates, or when, instead of likelihoods, Bayesian predictive densities are used [12, 13].

When N is small, the two decomposition terms on the right in (13) need not have similar values. In the maximum likelihood context, they have distinct and interesting interpretations. Maximization results in a larger value $L_N(\hat{\boldsymbol{\theta}}_N)$ than the ideal $L_N(\boldsymbol{\theta}_0)$, so the difference $\ln L_N(\hat{\boldsymbol{\theta}}_N) - \ln L_N(\boldsymbol{\theta}_0)$ quantifies the *overfit* of the model to the observed data due to parameter estimation. Similarly, the use of $L_N^*(\hat{\boldsymbol{\theta}}_N)$ with independent replicates instead of $L_N^*(\boldsymbol{\theta}_0)$, which maximizes $E_{X_1^*,\ldots,X_N^*}[\ln L_N^*(\boldsymbol{\theta})]$, results in an increase in Kullback information* discrepancy from the true density in the amount

$$E_{X_1^*,\ldots,X_N^*}[\ln L_N^*(\boldsymbol{\theta}_0)] - E_{X_1^*,\ldots,X_N^*}[\ln L_N^*(\hat{\boldsymbol{\theta}}_N)].$$

Hence this quantity measures the *accuracy loss* due to parameter estimation. The asymptotic equality of the decomposition components in (13), which does not require correct model assumptions, can be expressed as a connection between overfit and accuracy loss,

$$\text{mean overfit} \approx \text{mean accuracy loss}.$$

(In ref. [6], this result is called an *overfitting principle*.)

Thus, in many ways, Akaike's approach to the definition of AIC illuminates fundamental issues of statistical modeling.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, B. N. Petrov and F. Czaki, eds. Akadémia Kiadó, Budapest, pp. 267–281. Reproduced with an introduction by J. deLeeuw in *Breakthroughs in Statistics 1*, S. Kotz and N. L. Johnson, eds., Springer-Verlag, New York, pp. 599–624. (Derivation of AIC from fundamental principles with some applications, mainly to spectrum estimation.)

[2] Akaike, H. (1980). Likelihood and Bayes procedure. In *Bayesian Statistics*, J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds. University Press, Valencia, pp. 143–166. (Variety of applications of AIC and of a generalization of AIC for certain Bayesian models.)

[3] Akaike, H. (1985). Prediction and entropy. In *A Celebration of Statistics*, A. C. Atkinson and S. E. Fienberg, eds. Springer-Verlag, New York, pp. 1–24. (ISI Centenary volume.) Reprinted (1998) in *Selected Papers of Hirotugu Akaike*, E. Parzen et al., eds. Springer-Verlag, New York, pp. 387–410. (An expository paper; it provides an explanation of the appearance of the number 2 in some model selection contexts, derivations and reinterpretations of Kullback information and of some Bayesian principles, and a derivation of AIC. It discusses the difference between model selection by MinAIC and by hypothesis testing.)

[4] Bhansali, R. J. (1988). Consistent order determination for processes with infinite variance. *J. Roy. Statist. Soc. B*, **50**, 46–60. (Consistency of the minimum AIC order selection for finite-order autoregressive processes with infinite variance.)

[5] Findley, D. F. (1985). On the unbiasedness property of AIC for exact or approximating linear stochastic time series models. *J. Time Ser. Anal.*, **6**, 229–252. [Derivation of limit of left-hand side of (12) for not necessarily correct autoregressive moving-average time series models.]

[6] Findley, D. F. (1990). Counterexamples to parsimony and BIC. *Ann. Inst. Statist. Math.*, **43**, 505–514. (Examples of poor performance of "consistent" order selection criteria with incorrect models. Theoretical perspective on "overfitting.")

[7] Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C., Chen, B.-C. (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *J. Bus. Econ. Statist.*, **16**, 127–177, with discussion. (Description of Census Bureau's new time series modeling and seasonal adjustment program with many model comparison diagnostics.)

[8] Findley, D. F. and Parzen, E. (1995). A conversation with Hirotugu Akaike. *Statist. Sci.*, **10**, 104–117. (Interview with Akaike.)

[9] Fujikoshi, Y. and Satoh, K. (1997). Modified *AIC* and C_P in multivariate linear regression. *Biometrika*, **84**, 707–716. (Finite-sample variant

MAIC of AIC for both underparametrized and overparametrized linear regression models.)

[10] Galbraith, R. F. and Galbraith, J. I. (1974). On the inverses of some patterned matrices arising in the theory of stationary time series. *J. Appl. Probab.*, **11**, 63–71.

[11] Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307. (Small-sample version AIC_C of AIC for Gaussian linear regressions and autoregressions.)

[12] Ishiguro, M., Sakamoto, Y., and Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Ann. Inst. Statist. Math.*, **49**, 411–434. [Generalization EIC of AIC in which the left-hand of (12) is estimated for fixed N and non-MLE parameter estimates via the bootstrap.]

[13] Konishi, S. and Kitagawa, G. (1996). Generalized information criteria in model selection. *Biometrika*, **83**, 875–890. (Generalization GIC of AIC for models whose parameter estimates are statistical functionals* that need not be maximum likelihood estimates, including some robustly estimated models and approximate Bayesian predictive densities.)

[14] Lai, T.-L. and Wei, C.-Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist.*, **10**, 154–166.

[15] Linhart, H. and Zucchini, W. (1986). *Model Selection*. Wiley, New York. (Development of analogues of AIC for different estimation criteria with some applications.)

[16] Pötscher, B. M. (1985). The behavior of the Lagrangian multiplier test in testing the orders of an ARMA model. *Metrika*, **32**, 129–150. (Thorough discussion of subtleties of deriving model comparison test distributions for autoregressive moving-average time series models.)

[17] Sakamoto, Y. (1991). *Categorical Data Analysis by AIC*. Kluwer, Dordrecht.

[18] Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1985). *Akaike Information Criterion Statistics*. Reidel, Dordrecht. (Intermediate-level textbook approaching various basic statistical problems as model comparison problems.)

[19] Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, **63**, 117–126. (Asymptotic probabilities of overparametrized choices by the minimum AIC criterion.)

[20] Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.*, **8**, 147–164. (Optimality property of the minimum AIC criterion with autoregressions.)

[21] Shibata, R. (1989). Statistical aspects of model selection. In *From Data to Model*, J. C. Willems, ed. Springer, Berlin, pp. 215–240. (Derivations of generalizations TIC and RIC of AIC and a proof of the equivalence of a cross-validation* criterion with Takeuchi's TIC.)

[22] Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm. Statist. A*, **7**, 13–26. (Small-sample version of AIC for Gaussian linear regressions.)

(AKAIKE'S CRITERION
GENERALIZED FINAL PREDICTION ERROR CRITERIA
KULLBACK INFORMATION
LINEAR MODEL SELECTION, CRITERIA AND TESTS
MAXIMUM LIKELIHOOD ESTIMATION
REGRESSION VARIABLES, SELECTION OF
SCHWARZ CRITERION
STATISTICAL MODELING)

DAVID F. FINDLEY

ARCHAEOLOGY, STATISTICS IN (UPDATE)

Applications of statistics to archaeological data interpretation are widespread and can be divided broadly into two groups: those which are descriptive in nature (used primarily to reduce large and/or complex data sets to a more manageable size) and those which are model-based (used to make inferences about the underlying processes that gave rise to the data we observe). Approaches of the first type are most commonly adopted and, in general, are appropriately used and well understood by members of the archaeological profession. Model-based approaches are less widely used and usually rely upon collaboration with a professional statistician.

In ESS vol. 1 Gelfand provided an excellent survey of the application of statistics to archaeology up to and including the late 1970s. This entry supplements the earlier one, and the emphasis is on work undertaken since that time. Even so, this entry is not exhaustive, and readers are also encouraged to consult the review article of Fieller [15]. Statistics forms an increasingly important part of both undergraduate