ELSEVIER

Statistics & Probability Letters 33 (1997) 201–208

STATISTICS & PROBABILITY LETTERS

Unifying the derivations for the Akaike and corrected Akaike information criteria

Joseph E. Cavanaugh

Department of Statistics, University of Missouri, 222 Math Sciences Bldg., Columbia, MO 65211, USA

Received January 1996; revised May 1996

Abstract

The Akaike (1973, 1974) information criterion, AIC, and the corrected Akaike information criterion (Hurvich and Tsai, 1989), AICc, were both designed as estimators of the expected Kullback–Leibler discrepancy between the model generating the data and a fitted candidate model. AIC is justified in a very general framework, and as a result, offers a crude estimator of the expected discrepancy: one which exhibits a potentially high degree of negative bias in small-sample applications (Hurvich and Tsai, 1989). AICc corrects for this bias, but is less broadly applicable than AIC since its justification depends upon the form of the candidate model (Hurvich and Tsai, 1989, 1993; Hurvich et al., 1990; Bedrick and Tsai, 1994).

Although AIC and AICc share the same objective, the derivations of the criteria proceed along very different lines, making it difficult to reconcile how AICc improves upon the approximations leading to AIC. To address this issue, we present a derivation which unifies the justifications of AIC and AICc in the linear regression framework.

Keywords: AIC; AICc; Information theory; Kullback–Leibler information; Model selection

1. Introduction

The Akaike information criterion, AIC, was developed by Akaike (1973, 1974) to estimate the expected Kullback–Leibler discrepancy between the model generating the data and a fitted candidate model. In instances where the sample size is large and the dimension of the candidate model is relatively small, AIC serves as an approximately unbiased estimator. In other settings, AIC may be characterized by a large negative bias (Hurvich and Tsai, 1989) which limits its effectiveness as a model selection criterion. For such instances, Hurvich and Tsai (1989) proposed the corrected Akaike information criterion, AICc.

Originally suggested for linear regression by Sugiura (1978), Hurvich and Tsai (1989) justified AICc for linear and nonlinear regression and autoregressive modeling, and investigated the small-sample superiority of AICc over AIC in these settings. Since then, AICc has been extended to a number of additional frameworks, including autoregressive moving-average modeling (Hurvich et al., 1990), vector autoregressive modeling (Hurvich and Tsai, 1993), and multivariate regression modeling (Bedrick and Tsai, 1994).

PII S0167-7152(96)00128-9

The advantage of AICc over AIC is that in small-sample applications, AICc estimates the expected discrepancy with less bias than AIC. The advantage of AIC over AICc is that AIC is more universally applicable, since the derivation of AIC is quite general whereas the derivation of AICc relies upon the form of the candidate model.

Although AIC and AICc share the same fundamental objective, the justifications of the criteria provided by Akaike (1973) and Hurvich and Tsai (1989) proceed along different directions, making it difficult to reconcile how AICc refines the approximations used to establish AIC. A derivation which links the justifications would be instructive. With this as our goal, we formulate a derivation which connects the motivations for the criteria in the setting of the linear regression model. We consider this framework in order to keep our development straightforward: analogous derivations could be easily constructed for other settings in which AICc has been justified. Our work leads to some useful guidelines on characterizing linear regression settings in which AIC may provide an inadequate estimator of the expected discrepancy.

2. A unified derivation of AIC and AICc

A well-known measure of separation between two models is given by the nonnormalized Kullback–Leibler information (Kullback, 1968), also known as the cross entropy or discrepancy. Let θ_0 represent the set of parameters for the "true" or generating model, and let θ represent the set of parameters for an approximating or candidate model. Let Θ represent the parameter space for θ. The discrepancy between the generating model and the candidate model is defined as

$$d_n(\theta, \theta_0) = E_0\{-2 \ln L(\theta \mid Y_n)\},$$

where E_0 denotes the expectation under the generating model, and $L(\theta \mid Y_n)$ represents the likelihood corresponding to the candidate model.

Now for a given set of maximum likelihood estimates $\hat{\theta}_n$,

$$d_n(\hat{\theta}_n, \theta_0) = E_0\{-2 \ln L(\theta \mid Y_n)\}|_{\theta=\hat{\theta}_n} \tag{2.1}$$

would provide a useful measure of the separation between the generating model and the fitted candidate model. Yet evaluating (2.1) is not possible, since doing so requires knowledge of θ_0. Akaike (1973), however, noted that $-2 \ln L(\hat{\theta}_n \mid Y_n)$ serves as a biased estimator of (2.1), and that the bias adjustment

$$E_0\{E_0\{-2 \ln L(\theta \mid Y_n)\}|_{\theta=\hat{\theta}_n}\} - E_0\{-2 \ln L(\hat{\theta}_n \mid Y_n)\} \tag{2.2}$$

can often be asymptotically estimated by twice the dimension of $\hat{\theta}_n$.

Thus, if we let k represent the dimension of $\hat{\theta}_n$, then under appropriate conditions, the expected value of

$$\text{AIC} = -2 \ln L(\hat{\theta}_n \mid Y_n) + 2k \tag{2.3}$$

should be asymptotically near the expected value of (2.1), say

$$\Delta_n(k, \theta_0) = E_0\{d_n(\hat{\theta}_n, \theta_0)\}.$$

Specifically, one can establish that

$$E_0\{\text{AIC}\} + \text{o}(1) = \Delta_n(k, \theta_0). \tag{2.4}$$

The demonstration of this fact requires the strong assumption that θ_0 is an element of Θ; i.e., that the candidate family includes the generating model (see Linhart and Zucchini, 1986, p. 245).

AIC provides us with an approximately unbiased estimator of $\Delta_n(k, \theta_0)$ in settings where n is large and k is comparatively small. Yet in other settings, $2k$ may be much smaller than the bias adjustment (2.2), making AIC substantially negatively biased as an estimator of $\Delta_n(k, \theta_0)$. To correct for this negative bias, Hurvich and Tsai (1989) proposed AICc for linear and nonlinear regression and autoregressive modeling. For simplicity, we introduce and consider AICc in the context of the linear regression model, although our development here could easily be extended to other settings in which AICc has been justified.

Suppose that the generating model for the data is given by

$$y = X\beta_0 + e, \quad e \sim \mathrm{N}_n(0, \sigma_0^2 I), \tag{2.5}$$

and that the candidate model postulated for the data is of the form

$$y = X\beta + e, \quad e \sim \mathrm{N}_n(0, \sigma^2 I). \tag{2.6}$$

Here, y is an $n \times 1$ observation vector, e is an $n \times 1$ error vector, β_0 and β are $p \times 1$ parameter vectors, and X is an $n \times p$ design matrix of full-column rank.

Now, assume β_0 is such that for some $0 < p_0 \leqslant p$, the last $(p - p_0)$ components of β_0 are zero. Thus, model (2.5) is nested within model (2.6). Let θ_0 and θ respectively denote the $k = (p+1)$-dimensional vectors $(\beta'_0, \sigma_0^2)'$ and $(\beta', \sigma^2)'$. Note that the nesting ensures that θ_0 is an element of Θ, which is required to prove the analogue of (2.4) for AICc (see Hurvich and Tsai, 1989, p. 299).

Let $\hat{\beta}_n$ denote the least-squares estimator of β, and let $\hat{\sigma}_n^2 = (y - X\hat{\beta}_n)'(y - X\hat{\beta}_n)/n$. Hurvich and Tsai (1989, p. 300) define AICc as

$$n \ln \hat{\sigma}_n^2 + \frac{n(n+p)}{n-p-2}. \tag{2.7}$$

For convenience, we will use the operationally equivalent definition

$$\mathrm{AICc} = n \ln \hat{\sigma}_n^2 + n \ln 2\pi + \frac{n(n+p)}{n-p-2} = \{n \ln \hat{\sigma}_n^2 + n(1 + \ln 2\pi)\} + \frac{2(p+1)n}{n-p-2}, \tag{2.8}$$

which differs from (2.7) by an additive constant that has no impact on the selection behavior of the criterion.

One can prove that in the linear regression setting,

$$E_0\{\mathrm{AICc}\} = \Delta_n(k, \theta_0), \tag{2.9}$$

thus establishing that AICc is exactly unbiased for $\Delta_n(k, \theta_0)$. (The preceding holds up to o(1) for other modeling frameworks in which AICc has been justified and developed.)

We will derive AIC and verify (2.4) in a general setting, and derive AICc and verify (2.9) in the linear regression setting. Our objective will be to do so in a manner which will clearly illustrate the way in which AICc improves upon the approximations leading to AIC.

To begin, consider writing $\Delta_n(k, \theta_0)$ as follows:

$$\begin{aligned} \Delta_n(k, \theta_0) &= E_0\{E_0\{-2 \ \ln \ L(\theta \mid Y_n)\}|_{\theta = \hat{\theta}_n}\} \\ &= E_0\{-2 \ \ln \ L(\hat{\theta}_n \mid Y_n)\} && (2.10) \\ &\quad + [E_0\{-2 \ \ln \ L(\theta_0 \mid Y_n)\} - E_0\{-2 \ \ln \ L(\hat{\theta}_n \mid Y_n)\}] && (2.11) \\ &\quad + [E_0\{E_0\{-2 \ \ln \ L(\theta \mid Y_n)\}|_{\theta = \hat{\theta}_n}\} - E_0\{-2 \ \ln \ L(\theta_0 \mid Y_n)\}]. && (2.12) \end{aligned}$$

The derivation of AIC and verification of (2.4) are accomplished by establishing the following lemma, which asserts that (2.11) and (2.12) are both within o(1) of k. The lemma can be justified as a special case of Propositions 1 and 2 of Linhart and Zucchini (1986, pp. 240–242). A brief proof is sketched here for completeness.

We assume the necessary regularity conditions required to ensure the consistency and asymptotic normality of the maximum likelihood vector $\hat{\theta}_n$.

Lemma 1.

$$E_0\{-2 \ln L(\theta_0 \mid Y_n)\} - E_0\{-2 \ln L(\hat{\theta}_n \mid Y_n)\} = k + o(1), \tag{2.13}$$

$$E_0\{E_0\{-2 \ln L(\theta \mid Y_n)\}|_{\theta=\hat{\theta}_n}\} - E_0\{-2 \ln L(\theta_0 \mid Y_n)\} = k + o(1). \tag{2.14}$$

Proof. Define

$$I_n(\theta,\theta_0) = E_0\left[-\frac{\partial^2 \ln L(\theta \mid Y_n)}{\partial\theta\,\partial\theta'}\right] \quad \text{and} \quad \mathscr{I}_n(\theta, Y_n) = \left[-\frac{\partial^2 \ln L(\theta \mid Y_n)}{\partial\theta\,\partial\theta'}\right].$$

First, consider taking a second-order expansion of $-2\ln L(\theta_0 \mid Y_n)$ about $\hat{\theta}_n$ and evaluating the expectation (under θ_0) of the result. We obtain

$$\begin{aligned} E_0\{-2\ln L(\theta_0 \mid Y_n)\} &= E_0\{-2\ln L(\hat{\theta}_n \mid Y_n)\} \\ &\quad + E_0\{(\hat{\theta}_n - \theta_0)'\{\mathscr{I}_n(\hat{\theta}_n, Y_n)\}(\hat{\theta}_n - \theta_0)\} + o(1). \end{aligned} \tag{2.15}$$

Next, consider taking a second-order expansion of $E_0\{-2\ln L(\theta \mid Y_n)\}|_{\theta=\hat{\theta}_n}$ about θ_0, again evaluating the expectation (under θ_0) of the result. We obtain

$$\begin{aligned} E_0\{E_0\{-2\ln L(\theta \mid Y_n)\}|_{\theta=\hat{\theta}_n}\} &= E_0\{-2\ln L(\theta_0 \mid Y_n)\} \\ &\quad + E_0\{(\hat{\theta}_n - \theta_0)'\{I_n(\theta_0,\theta_0)\}(\hat{\theta}_n - \theta_0)\} + o(1). \end{aligned} \tag{2.16}$$

Now the quadratic forms

$$(\hat{\theta}_n - \theta_0)'\{\mathscr{I}_n(\hat{\theta}_n, Y_n)\}(\hat{\theta}_n - \theta_0) \quad \text{and} \quad (\hat{\theta}_n - \theta_0)'\{I_n(\theta_0,\theta_0)\}(\hat{\theta}_n - \theta_0)$$

both converge to centrally distributed chi-square random variables with k degrees of freedom. (Recall again that we are assuming $\theta_0 \in \Theta$.) Thus, the expectations (under θ_0) of both quadratic forms are within o(1) of k. This fact along with (2.15) and (2.16) establishes (2.13) and (2.14). □

The asymptotic approximation of the sum $\{(2.11) + (2.12)\}$ by $2k$ may be somewhat crude for small or moderate sample-size applications. However, for a more precise assessment of this quantity, one would need to further specify the underlying modeling framework. For example, in the linear regression setting of interest, the following lemma will show that $\{(2.11) + (2.12)\}$ can be exactly evaluated as a function of $p = (k - 1)$ and n. This evaluation will lead to the derivation of AICc and verification of (2.9).

Lemma 2. *For the generating model* (2.5) *and the candidate model* (2.6),

$$E_0\{-2 \ln L(\theta_0 \mid Y_n)\} - E_0\{-2 \ln L(\hat{\theta}_n \mid Y_n)\} = -E_0\left\{n \ln \frac{\hat{\sigma}_n^2}{\sigma_0^2}\right\}, \tag{2.17}$$

$$E_0\{E_0\{-2 \ln L(\theta \mid Y_n)\}|_{\theta=\hat{\theta}_n}\} - E_0\{-2 \ln L(\theta_0 \mid Y_n)\} = E_0\left\{n \ln \frac{\hat{\sigma}_n^2}{\sigma_0^2}\right\} + \frac{2(p+1)n}{n-p-2}. \tag{2.18}$$

Proof. The log-likelihood for the candidate model (2.6) is given by

$$\ln L(\theta \mid Y_n) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta).$$

Under the generating model (2.5), one can easily establish the following relations:

$$E_0\{-2\ \ln\ L(\theta_0 \mid Y_n)\} = n\ln\sigma_0^2 + n(1+\ln 2\pi), \tag{2.19}$$

$$E_0\{-2\ \ln\ L(\hat{\theta}_n \mid Y_n)\} = E_0\{n\ln\hat{\sigma}_n^2\} + n(1+\ln 2\pi), \tag{2.20}$$

$$E_0\{-2\ \ln\ L(\theta \mid Y_n)\}|_{\theta=\hat{\theta}_n} = n\ln\hat{\sigma}_n^2 + \frac{n\sigma_0^2}{\hat{\sigma}_n^2} + \frac{1}{\hat{\sigma}_n^2}(\hat{\beta}_n-\beta_0)'(X'X)(\hat{\beta}_n-\beta_0) + n\ln 2\pi. \tag{2.21}$$

Now to evaluate the expected value of (2.21) under (2.5), note that $(n\hat{\sigma}_n^2/\sigma_0^2)$ has a chi-square distribution with $(n-p)$ degrees of freedom, the quadratic form $\{(\hat{\beta}_n-\beta_0)'\{(1/\sigma_0^2)(X'X)\}(\hat{\beta}_n-\beta_0)\}$ has a chi-square distribution with p degrees of freedom, and $\hat{\sigma}_n^2$ and $\hat{\beta}_n$ are independent. Using the fact that the expectation of the reciprocal of a chi-square random variable with df degrees of freedom is given by $(1/(\mathrm{df}-2))$, one can argue that the expectation of (2.21) under (2.5) reduces to

$$E_0\{E_0\{-2\ \ln\ L(\theta \mid Y_n)\}|_{\theta=\hat{\theta}_n}\} = E_0\{n\ln\hat{\sigma}_n^2\} + \frac{n^2}{n-p-2} + \frac{np}{n-p-2} + n\ln 2\pi \tag{2.22}$$

(see Hurvich and Tsai, 1989, p. 299).

By (2.19) and (2.20), we have (2.17). By (2.19) and (2.22), we have (2.18). □

AICc is derived and (2.9) verified by using Lemma 2 in conjunction with the representation of $\Delta_n(k,\theta_0)$ as the sum of (2.10)–(2.12). Note that in the present setting, we can use (2.17) and (2.18) to write $\Delta_n(k,\theta_0)$ as follows:

$$\begin{aligned}\Delta_n(k,\theta_0) &= E_0\{E_0\{-2\ \ln\ L(\theta \mid Y_n)\}|_{\theta=\hat{\theta}_n}\}\\ &= E_0\{-2\ \ln\ L(\hat{\theta}_n \mid Y_n)\}\\ &\quad - E_0\left\{n\ln\frac{\hat{\sigma}_n^2}{\sigma_0^2}\right\} + E_0\left\{n\ln\frac{\hat{\sigma}_n^2}{\sigma_0^2}\right\} + \frac{2(p+1)n}{n-p-2}\\ &= E_0\{-2\ \ln\ L(\hat{\theta}_n \mid Y_n)\} + \frac{2(p+1)n}{n-p-2}.\end{aligned}$$

The preceding justifies that

$$-2\ \ln\ L(\hat{\theta}_n \mid Y_n) + \frac{2(p+1)n}{n-p-2} \tag{2.23}$$

is exactly unbiased for $\Delta_n(k,\theta_0)$. Yet since

$$-2\ln L(\hat{\theta}_n \mid Y_n) = n\ln\hat{\sigma}_n^2 + n(1+\ln 2\pi),$$

(2.23) is precisely the same as the definition of AICc provided by (2.8).

In summary, we have derived AIC and verified (2.4) through utilizing Lemma 1 along with the representation of $\Delta_n(k,\theta_0)$ as the sum of (2.10)–(2.12). For the linear regression setting, we have derived AICc and verified (2.9) through utilizing Lemma 2 along with the same representation of $\Delta_n(k,\theta_0)$.

The third and final lemma provides a predictable albeit meaningful connection between AIC and AICc.

Lemma 3. *For the generating model* (2.5) *and the candidate model* (2.6),

$$-E_0\left\{n\ln\frac{\hat{\sigma}_n^2}{\sigma_0^2}\right\} = (p+1) + \mathrm{o}(1), \tag{2.24}$$

$$E_0\left\{n\ln\frac{\hat{\sigma}_n^2}{\sigma_0^2}\right\} + \frac{2(p+1)n}{n-p-2} = (p+1) + \mathrm{o}(1). \tag{2.25}$$

Proof. We begin by considering an asymptotic approximation for $-E_0\{n\ln(\hat{\sigma}_n^2/\sigma_0^2)\}$.

Let X_{df} represent a chi-square random variable with df degrees of freedom. By taking a second-order expansion of $\ln X_{\mathrm{df}}$ about df and evaluating the expectation of the result, one can justify the relation

$$E\{\ln X_{\mathrm{df}}\} = \ln \mathrm{df} - \frac{1}{\mathrm{df}} + \mathrm{O}\left(\frac{1}{\mathrm{df}^2}\right)$$

(see Bickel and Doksum, 1977, p. 31). Using this relation along with the fact that $(n\hat{\sigma}_n^2/\sigma_0^2)$ has a chi-square distribution with $(n-p)$ degrees of freedom, we can write

$$\begin{aligned}-E_0\left\{n\ln\frac{\hat{\sigma}_n^2}{\sigma_0^2}\right\} &= n\ln n - nE_0\left\{\ln\frac{n\hat{\sigma}_n^2}{\sigma_0^2}\right\}\\ &= n\ln n - n\left\{\ln(n-p) - \frac{1}{n-p} + \mathrm{O}\left(\frac{1}{(n-p)^2}\right)\right\}.\end{aligned} \tag{2.26}$$

Next, consider taking a first-order expansion of $\ln(n-p)$ about n to obtain

$$\ln(n-p) = \ln n - \frac{p}{n} + \mathrm{O}\left(\frac{p^2}{n^2}\right). \tag{2.27}$$

Substituting (2.27) into (2.26) yields

$$-E_0\left\{n\ln\frac{\hat{\sigma}_n^2}{\sigma_0^2}\right\} = (p+1) + \frac{p}{n-p} + \mathrm{O}\left(\frac{p^2}{n}\right) + \mathrm{O}\left(\frac{n}{(n-p)^2}\right). \tag{2.28}$$

Asymptotically, each of the final three terms on the right-hand side of (2.28) is $\mathrm{o}(1)$ as $n \to \infty$ when p is held constant. However, each of these terms is also $\mathrm{o}(1)$ when p is allowed to grow at a rate less than $\sqrt{n}$ as $n \to \infty$. Thus, assuming $n \to \infty$ and p is $\mathrm{O}(n^{(1/2)-\delta})$, $\delta > 0$, (2.28) allows us to write

$$-E_0\left\{n\ln\frac{\hat{\sigma}_n^2}{\sigma_0^2}\right\} = (p+1) + \mathrm{o}(1).$$

This establishes (2.24). The justification of (2.25) follows since we can write

$$\begin{aligned}E_0\left\{n\ln\frac{\hat{\sigma}_n^2}{\sigma_0^2}\right\} + \frac{2(p+1)n}{n-p-2} &= \{-(p+1) + \mathrm{o}(1)\} + \left\{2(p+1) + \frac{2(p+1)(p+2)}{n-p-2}\right\}\\ &= (p+1) + \mathrm{o}(1). \quad \square\end{aligned}$$

3. Discussion

The development in Section 2 along with the results of Hurvich and Tsai (1989) suggest that for linear regression, AICc is preferable to AIC as an estimator of $\Delta_n(k,\theta_0)$ unless p is "small" and n is "large". The penalty term of AICc is exactly equal to the sum $\{(2.11) + (2.12)\}$, whereas the penalty term of AIC only provides an approximation to this sum. How crude is the approximation of (2.11) and (2.12) by $k = (p + 1)$? To address this question in the linear regression setting, we evaluate (2.11) and (2.12) for various values of p and n and present the results in Table 1. The exact computation of (2.11) and (2.12) is possible using the fact that

$$E_0\left\{\ln \frac{n\hat{\sigma}_n^2}{\sigma_0^2}\right\} = \ln 2 + \psi\left(\frac{n-p}{2}\right),$$

where ψ is the *digamma* or *psi* function (Kotz et al., 1982, p. 373) (see Hurvich and Tsai, 1989, p. 303). Bernardo (1976) presents a simple algorithm for computing values of ψ.

The results help to illustrate two interesting principles. First, the approximation of (2.12) by $(p + 1)$ is uniformly worse than the approximation of (2.11) by $(p + 1)$, due to the fact that $(p + 1)$ is always less than (2.11) and (2.11) is always less than (2.12). Thus, the bias correction provided by AICc to AIC plays a more important role in refining the approximation of (2.12) than in refining the approximation of (2.11), particularly where n is small and p is large.

Second, in reference to (2.24) and (2.25), it can be easily shown that if $n > (p + 2)(2p + 3)$, then $2(p + 1)$ is less than one unit smaller than $(2(p + 1)n)/(n - p - 2)$. If n exceeds $(p + 2)(2p + 3)$, then $(p + 1)$ appears to provide a good approximation to (2.12), and hence to (2.11). However, if n is substantially less than $(p + 2)(2p + 3)$, then $(p + 1)$ may be much smaller than (2.12) (or (2.11)) even when n is "large". In such settings, AIC may greatly underestimate $\Delta_n(k, \theta_0)$, and as a result, may perform poorly as a model selection criterion.

We close by noting that the development in Section 2 suggests that $-2 \ln L(\hat{\theta}_n \mid Y_n)$ is a justifiable choice for the goodness-of-fit term in AICc. The derivations of AICc presented by Hurvich and Tsai (1989, 1993) and

Table 1
Evaluation of (2.11) and (2.12) for various values of n and p

p	n	(2.11)	(2.12)	p	n	(2.11)	(2.12)
2	20	3.24	4.26	2	80	3.06	3.26
4	20	5.74	8.55	4	80	5.16	5.65
8	20	11.93	24.07	8	80	9.55	11.03
16	20	37.60	302.40	16	80	19.11	24.76
2	30	3.15	3.77	2	160	3.03	3.13
4	30	5.46	7.04	4	160	5.08	5.31
8	30	10.69	16.31	8	160	9.26	9.94
16	30	25.06	59.94	16	160	17.97	20.34
2	40	3.11	3.55	2	320	3.01	3.06
4	40	5.34	6.43	4	320	5.04	5.15
8	40	10.19	13.81	8	320	9.13	9.45
16	40	22.12	39.70	16	320	17.46	18.56
2	50	3.09	3.43	2	640	3.01	3.03
4	50	5.26	6.10	4	640	5.02	5.08
8	50	9.92	12.58	8	640	9.06	9.22
16	50	20.77	32.36	16	640	17.23	17.75

$(2.11) = E_0\{-2 \ln L(\theta_0 \mid Y_n)\} - E_0\{-2 \ln L(\hat{\theta}_n \mid Y_n)\}$.
$(2.12) = E_0\{E_0\{-2 \ln L(\theta \mid Y_n)\}|_{\theta = \hat{\theta}_n}\} - E_0\{-2 \ln L(\theta_0 \mid Y_n)\}$.

Bedrick and Tsai (1994) arrive at goodness-of-fit terms of the form $n \ln \hat{\sigma}_n^2$, where $\hat{\sigma}_n^2$ represents the estimate of the error variance. Of course in linear regression, the terms $n \ln \hat{\sigma}_n^2$ and $-2 \ln L(\hat{\theta}_n \mid Y_n)$ differ only by the additive constant $n(1 + \ln 2\pi)$; thus, it does not matter which quantity is used for the goodness-of-fit term. Yet in many applications, such as with autoregressive or autoregressive moving-average models, there may be a substantial difference between $n \ln \hat{\sigma}_n^2$ and $-2 \ln L(\hat{\theta}_n \mid Y_n)$. Hurvich et al. (1990) present a convincing case for preferring the latter to the former as a goodness-of-fit term in defining model selection criteria of the same basic form of AIC or AICc: e.g., the criteria of Schwarz (1978), Hannan and Quinn (1979), etc. The development in Section 2 shows that the choice of $-2 \ln L(\hat{\theta}_n \mid Y_n)$ in the definition of AICc is theoretically well motivated.

Acknowledgements

The author wishes to express his gratitude to an anonymous referee for carefully reading and constructively critiquing the original version of this manuscript.

References

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, in: B.N. Petrov and F. Csaki, eds., *2nd Internat. Symp. on Information Theory* (Akademia Kiado, Budapest) pp. 267–281.

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Automat. Control* **AC-19**, 716–723.

Bedrick, E.J. and C.L. Tsai (1994), Model selection for multivariate regression in small samples, *Biometrics* **50**, 226–231.

Bernardo, J.M. (1976), Psi (digamma) function, *Appl. Statist.* **25**, 315–317.

Bickel, P.J. and K.A. Doksum (1977), *Mathematical Statistics: Basic Ideas and Selected Topics* (Prentice-Hall, Englewood Cliffs, NJ).

Hannan, E.J. and B.G. Quinn (1979), The determination of the order of an autoregression, *J. Roy. Statist. Soc. B* **41**, 190–195.

Hurvich, C.M., R.H. Shumway and C.L. Tsai (1990), Improved estimators of Kullback–Leibler information for autoregressive model selection in small samples, *Biometrika* **77**, 709–719.

Hurvich, C.M. and C.L. Tsai (1989), Regression and time series model selection in small samples, *Biometrika* **76**, 297–307.

Hurvich, C.M. and C.L. Tsai (1993), A corrected Akaike information criterion for vector autoregressive model selection, *J. Time Series Anal.* **14**, 271–279.

Kotz, S., N.L. Johnson and C.B. Read, eds. (1982), *Encyclopedia of Statistical Sciences*, Vol. 2 (Wiley, New York).

Kullback, S. (1968), *Information Theory and Statistics* (Dover, New York).

Linhart, H. and W. Zucchini (1986), *Model Selection* (Wiley, New York).

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Statist.* **6**, 461–464.

Sugiura, N. (1978), Further analysis of the data by Akaike's information criterion and the finite corrections, *Comm. Statist.* **A7**, 13–26.