J. R. Statist. Soc. B (2002)
64, *Part* 2, *pp.* 237–252

Regression model selection—a residual likelihood approach

Peide Shi

Peking University, Beijing, People's Republic of China

and Chih-Ling Tsai

University of California at Davis, USA

[Received June 2000. Final revision November 2001]

Summary. We obtain the residual information criterion RIC, a selection criterion based on the residual log-likelihood, for regression models including classical regression models, Box–Cox transformation models, weighted regression models and regression models with autoregressive moving average errors. We show that RIC is a consistent criterion, and that simulation studies for each of the four models indicate that RIC provides better model order choices than the Akaike information criterion, corrected Akaike information criterion, final prediction error, C_p and R^2_{adj}, except when the sample size is small and the signal-to-noise ratio is weak. In this case, none of the criteria performs well. Monte Carlo results also show that RIC is superior to the consistent Bayesian information criterion BIC when the signal-to-noise ratio is not weak, and it is comparable with BIC when the signal-to-noise ratio is weak and the sample size is large.

Keywords: Akaike information criterion; Bayesian information criterion; Corrected Akaike information criterion; C_p; Residual information criterion; Residual likelihood

1. Introduction

Over the last three decades, many model selection criteria such as the Akaike information criterion AIC (Akaike, 1973), corrected Akaike information criterion AICC (Hurvich and Tsai, 1989), Bayesian information criterion BIC (Schwarz, 1978), final prediction error FPE (Akaike, 1970) and C_p (Mallows, 1973) have been proposed and studied for linear regression models with constant variance. All these criteria have two basic elements: the first is a function of the scaled parameter estimator, which measures the goodness of fit; the second is a function of the number of unknown parameters, which penalizes overfitting. Selection criteria are usually classified into one of two categories on the basis of their second element: efficient (e.g. AIC, AICC, FPE and C_p) or consistent (e.g. BIC). Efficient criteria select the best finite dimensional candidate model in large samples when the true model is of infinite dimension. Consistent criteria select the correct model with probability approaching 1 in large samples when the true model is of finite dimension. There is no general agreement on which category is preferable for practical applications. For example, Burnham and Anderson (1998) preferred to use AIC-type efficient criteria to analyse empirical data in the biological and social sciences and medicine. By contrast, consistent criteria are often preferred in the physical sciences and engineering. However, because both efficiency and consistency are asymptotic

Address for correspondence: Chih-Ling Tsai, Graduate School of Management, University of California at Davis, Davis, CA 95616-8609, USA.
E-mail: cltsai@ucdavis.edu

1369–7412/02/64237

concepts and practitioners may not know the true nature of the problem at hand, we prefer a criterion whose performance is superior to or comparable with the above known consistent and efficient criteria in finite samples, even when it is not known whether the assumptions for efficiency or consistency are met.

For linear regression models with constant variance, Akaike (1973) used the likelihood function, as well as the maximum likelihood estimator of the scaled parameter, as a basis for obtaining AIC. However, this estimator is biased, and so we need to consider other methods of estimation. One such alternative is the method of residual likelihood (also called the restricted likelihood or the marginal likelihood), introduced by Patterson and Thompson (1971) and widely used for regression models with a general covariance structure (see Harville (1974), Corbeil and Searle (1976), Tunnicliffe Wilson (1989), Verbyla (1993), Diggle *et al.* (1994) and Cheang and Reinsel (2000)). Two useful references on residual likelihood are McCullagh and Nelder (1989) and Verbyla (1990).

The aim of this paper is to apply the residual log-likelihood approach to obtain a residual information criterion RIC, which has both BIC's useful property of consistency and a small sample performance that is comparable with that of AICC. This criterion is based on the expected Kullback–Leibler information of the residual log-likelihood of the fitted model. In the next section, a generalized version of the linear regression model is formulated which includes classical regression, Box–Cox transformation (Box and Cox, 1964), heteroscedasticity and autoregressive moving average (ARMA) errors. Selection criteria are then derived for each model. In Section 3 we prove that RIC is a consistent criterion. In Section 4, we compare RIC with AIC, AICC, BIC, FPE, C_p and R^2_{adj}, and we see that RIC performs well except when the sample size is small and the signal-to-noise ratio is weak. Section 5 presents concluding remarks.

2. Derivation of the residual information criterion

2.1. Model structures

Consider the collection of candidate models

$$h(y, \lambda) = X\beta + e, \tag{2.1}$$

where $y = (y_1, \ldots, y_n)'$ and $h(y, \lambda) = (h(y_1, \lambda), \ldots, h(y_n, \lambda))'$; $h(y_i, \lambda) = (y_i^\lambda - 1)/\lambda$ when $\lambda \neq 0$; otherwise $h(y_i, \lambda) = \log(y_i)$ for $i = 1, \ldots, n$. Here X is an $n \times k$ matrix, β is a $k \times 1$ vector, $e = (e_1, \ldots, e_n)'$ has a multivariate normal distribution with mean 0 and variance $\sigma^2\, W(\theta)$, σ is an unknown scalar and θ is an $m \times 1$ unknown vector. Many models are special cases of model (2.1), e.g. the Box–Cox heteroscedasticity model proposed by Lahiri and Egy (1981), which takes non-linearity and heteroscedasticity into account simultaneously. By adopting the results of Verbyla (1990) or Diggle *et al.* (1994), section 4.5, we can obtain the residual log-likelihood for the candidate model:

$$\begin{aligned} L\{(\theta', \sigma^2); (y, \lambda)\} = & -\tfrac{1}{2}(n-k)\log(\sigma^2) - \tfrac{1}{2}\log|W| - \tfrac{1}{2}\log|X^{*\prime}X^*| \\ & -\tfrac{1}{2}h^*(y, \lambda)'(I - H^*)\, h^*(y, \lambda)/\sigma^2 + \log(J), \end{aligned} \tag{2.2}$$

where $h^*(y, \lambda) = W^{-1/2}\, h(y, \lambda)$, $X^* = W^{-1/2}X$, $H^* = X^*(X^{*\prime}X^*)^{-1}X^{*\prime}$, $W = W(\theta)$ and $J = (\Pi y_i)^{\lambda-1}$ is the Jacobian of the power transformation of the candidate model.

Suppose that the data to be considered are consistent with a particular model which constitutes the nearest representation of the true situation. Adopting Linhart and Zucchini's (1986) terminology, we call that model the operating model:

$$h(y, \lambda_0) = X_0\beta_0 + \varepsilon, \tag{2.3}$$

where $h(y, \lambda_0)$, X_0 and β_0 are defined as in equation (2.1) except that the dimension of X_0 and β_0 is k_0 rather than k. In addition, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ has a multivariate normal distribution with mean $E(\varepsilon) = 0$ and variance $E(\varepsilon\varepsilon') = \sigma_0^2\, W(\theta_0)$, σ_0 is an unknown scalar and θ_0 is an $m_0 \times 1$ unknown vector. Then the residual log-likelihood for the operating model is $L_0\{(\theta_0', \sigma_0^2); (y, \lambda_0)\}$, which has the same form as equation (2.2) except that k, λ, σ^2, W and X in equation (2.2) are replaced by k_0, λ_0, σ_0^2, W_0 and X_0 respectively. We next use the residual likelihood function to obtain a model selection criterion.

2.2. Residual information criterion

A useful measure of the discrepancy between the operating and candidate residual log-likelihood functions is the Kullback–Leibler information. It is given by

$$d(\lambda, \theta', \sigma^2) = E_0[-2\, L\{(\theta', \sigma^2); (y, \lambda)\} + 2\, L_0\{(\theta_0', \sigma_0^2); (y, \lambda_0)\}], \tag{2.4}$$

where E_0 denotes the expectation under the residual likelihood function of the operating model. When λ lies in some neighbourhood of λ_0, the residual log-likelihood function (2.2) can be approximated by

$$L[\{\theta'(\lambda_0), \sigma^2(\lambda_0)\}; (y, \lambda_0)] + (\lambda - \lambda_0)\, A\{y, \lambda_0, \theta'(\lambda_0), \sigma^2(\lambda_0)\}\{1 + O(\lambda - \lambda_0)\}, \tag{2.5}$$

where

$$A\{y, \lambda_0, \theta'(\lambda_0), \sigma^2(\lambda_0)\} = \sum_{i=1}^{n} \log(y_i) - h^{*(1)}(y, \lambda_0)'[I - H^*\{\theta'(\lambda_0)\}]\; h^*(y, \lambda_0)/\sigma^2(\lambda_0),$$

$h^{*(1)}(y, \lambda_0)$ is the partial derivative of $h^*(y, \lambda)$ with respect to λ evaluated at λ_0 and $\theta'(\lambda_0)$ and $\sigma^2(\lambda_0)$ are the covariance and scaled parameters of the candidate model when the transformation parameter is λ_0. Thus equation (2.4) can be approximated by

$$d\{\lambda, \theta'(\lambda_0), \sigma^2(\lambda_0)\} = d\{\lambda_0, \theta'(\lambda_0), \sigma^2(\lambda_0)\} - 2(\lambda - \lambda_0)(E_0[A\{y, \lambda_0, \theta'(\lambda_0), \sigma^2(\lambda_0)\}]), \tag{2.6}$$

where

$$\begin{aligned} d\{\lambda_0, \theta'(\lambda_0), \sigma^2(\lambda_0)\} &= E_0[-2\, L\{\theta'(\lambda_0), \sigma^2(\lambda_0); (y, \lambda_0)\} + 2\, L_0\{(\theta_0', \sigma_0^2); (y, \lambda_0)\}] \\ &= (n-k)\, \log\{\sigma^2(\lambda_0)\} + \log|W\{\theta'(\lambda_0)\}| + \log|X'W^{-1}\{\theta'(\lambda_0)\}X| \\ &\quad + \operatorname{tr}(\beta_0' X_0' W'^{-1/2}\{\theta'(\lambda_0)\}[I - H^*\{\theta'(\lambda_0)\}]W^{-1/2}\{\theta'(\lambda_0)\}X_0\beta_0) \\ &\quad + \operatorname{tr}([I - H^*\{\theta'(\lambda_0)\}]W^{1/2}\{\theta'(\lambda_0)\}W_0 W'^{1/2}\{\theta'(\lambda_0)\})\sigma_0^2/\sigma^2\{\theta'(\lambda_0)\} \end{aligned}$$

(omitting irrelevant terms).

A reasonable criterion for judging the quality of the candidate model in the light of the data is $E_0[d\{\hat{\lambda}, \hat{\theta}'(\lambda_0), \tilde{\sigma}^2(\lambda_0)\}]$, where $\hat{\lambda}$ and $\hat{\theta}(\lambda_0)$ are estimates of λ and $\theta(\lambda_0)$ respectively,

$$\tilde{\sigma}^2(\lambda_0) = \hat{h}^*(y, \lambda_0)'(I - \hat{H}^*)\, \hat{h}^*(y, \lambda_0)/(n-k),$$

$\hat{X}^* = \hat{W}^{-1/2}X$, $\hat{H}^* = \hat{X}^*(\hat{X}^{*\prime}\hat{X}^*)^{-1}\hat{X}^{*\prime}$, $\hat{h}^*(y, \lambda_0) = \hat{W}^{-1/2}\, h(y, \lambda_0)$ and $\hat{W} = W\{\hat{\theta}(\lambda_0)\}$.

From equation (2.6), we have

$$\begin{aligned} d\{\hat{\lambda}, \hat{\theta}'(\lambda_0), \tilde{\sigma}^2(\lambda_0)\} = {} & (n-k)\log\{\tilde{\sigma}^2(\lambda_0)\} + \log|\hat{W}| + \log|\hat{X}^{*\prime}\hat{X}^*| \\ & + \operatorname{tr}\{\beta_0' X_0' \hat{W}'^{-1/2}(I-\hat{H}^*)\hat{W}^{-1/2}X_0\beta_0\} \\ & + \operatorname{tr}\{(I-\hat{H}^*)\hat{W}^{-1/2}W_0\hat{W}'^{-1/2}\}\sigma_0^2/\tilde{\sigma}^2(\lambda_0) - 2(\hat{\lambda}-\lambda_0)\hat{A}, \end{aligned} \tag{2.7}$$

where $\hat{A}$ is $E_0[A\{y, \lambda_0, \theta'(\lambda_0), \sigma^2(\lambda_0)\}]$ evaluated at $\theta(\lambda_0) = \hat{\theta}(\lambda_0)$ and $\sigma^2(\lambda_0) = \tilde{\sigma}^2(\lambda_0)$. We now assume that the set of candidate models includes the operating model, an assumption that is also used in the derivation of the Akaike information criterion AIC (Linhart and Zucchini, 1986) and the corrected Akaike information criterion AICC (Hurvich and Tsai, 1989). Under this assumption, the columns of X can be rearranged so that $X_0\beta_0 = X\beta^*$, where $\beta^* = (\beta_0', 0')'$ and $0'$ is a $1 \times (k-k_0)$ vector. Hence, the fourth term of the right-hand side of equation (2.7) is 0. Now suppose that, except for an event whose probability tends to 0 with n, the smallest eigenvalue of $\hat{X}^{*\prime}\hat{X}^*$ is greater than a positive number when n is sufficiently large. Then $\log|\hat{X}^{*\prime}\hat{X}^*|$ can be approximated by $k\log(n)[1 + O_p\{1/\log(n)\}]$. Under the assumption that the set of candidate models includes the operating model, it is reasonable to assume that $\hat{\theta}(\lambda_0)$ is a consistent estimate of θ_0, and therefore we can estimate W_0 by $\hat{W}$ (i.e. $W_0 = \hat{W} + o_p(1)$). If we replace W_0 by its approximation $\hat{W}$ in the second trace component of equation (2.7), then $d\{\hat{\lambda}, \hat{\theta}'(\lambda_0), \tilde{\sigma}^2(\lambda_0)\}$ can be approximated by

$$(n-k)\ \log\{\tilde{\sigma}^2(\lambda_0)\} + \log|\hat{W}| + k\ \log(n) + (n-k)\sigma_0^2/\tilde{\sigma}^2(\lambda_0) - 2(\hat{\lambda}-\lambda_0)\hat{A}. \tag{2.8}$$

It should be stressed that, since the above assumptions and approximations are made only with respect to the derivation of RIC, they do not impose any restrictions on our assessment of its performance.

Given a collection of candidate models, the one which minimizes $E_0[d\{\hat{\lambda}, \hat{\theta}'(\lambda_0), \tilde{\sigma}^2(\lambda_0)\}]$ is, in a sense, closest to the truth and is to be preferred. When the candidate models include the operating model, $(n-k)\ \tilde{\sigma}^2(\lambda_0)/\sigma_0^2$ is approximately distributed as χ^2_{n-k}, and

$$E_0\left\{\frac{\sigma_0^2}{\tilde{\sigma}^2(\lambda_0)}\right\} \approx \frac{n-k}{n-k-2}.$$

This, together with equation (2.8) (after subtracting $n+2$), gives us

$$\begin{aligned} E_0[d\{\hat{\lambda}, \hat{\theta}'(\lambda_0), \tilde{\sigma}^2(\lambda_0)\}] \approx {} & (n-k)\ E_0[\log\{\tilde{\sigma}^2(\lambda_0)\}] + E_0(\log|\hat{W}|) + k\{\log(n) - 1\} \\ & + \frac{4}{n-k-2} - 2\ E_0\{(\hat{\lambda}-\lambda_0)\hat{A}\}. \end{aligned} \tag{2.9}$$

In practice, the true transformation parameter is not known, and so we replace λ_0 in equation (2.9) by its estimator $\hat{\lambda}$. Consequently, an approximate unbiased estimate of equation (2.9) results in the residual information criterion

$$\mathrm{RIC} = (n-k)\ \log(\tilde{\sigma}^2) + \log|\hat{W}| + k\log(n) - k + \frac{4}{n-k-2}. \tag{2.10}$$

Both $\tilde{\sigma}^2$ and $\hat{W}$ are now evaluated at $\lambda = \hat{\lambda}$. Although there are several alternative criteria that can be derived from equation (2.9), we have chosen RIC because

(a) it is easy to compute,
(b) it has the useful asymptotic property of consistency (see Section 3) and
(c) it performs well in finite samples (see Section 4).

Two selection criteria in particular are often considered in model selection for classical regression models—BIC (which is consistent) and AIC (which is efficient). Therefore it will be useful to understand how RIC is similar to or different from BIC and AIC if it is to be of practical use. However, since Hurvich and Tsai (1989) showed that AICC outperforms AIC and other efficient criteria, we shall instead compare RIC with BIC and AICC. The third component of RIC and the second component of BIC $= n\log(\hat{\sigma}^2) + k\log(n)$ are identical, where $\hat{\sigma}^2$ is the maximum likelihood estimator of σ^2. In addition, the last component of RIC has the same denominator as the penalty function of AICC $= n\log(\hat{\sigma}^2) + 2n(k+1)/(n-k-2)$.

Next, we decompose the first component of RIC, BIC and AICC into two parts so that we can compare them effectively. The first parts of RIC and BIC are $(n-k)\log(\mathrm{SSE})$ and $n\log(\mathrm{SSE})$, and the second parts are $-(n-k)\log(n-k)$ and $-n\log(n)$ respectively, where SSE is the sum of squares of residuals. The first component of AICC has the same two parts as those of BIC. The first part measures the goodness of fit and restrains underfitting, whereas the second part together with the rest of the corresponding components of RIC, BIC and AICC are penalty functions for controlling overfitting.

It can be shown that RIC has a larger penalty function than BIC when $n > 2$. In the case where the signal-to-noise ratio is strong, the model is easily identified and overfitting is a problem. Here RIC's large penalty function allows it to perform better than BIC, and this expectation will be confirmed in our simulation results in Section 4. In addition, simulation studies show that RIC outperforms AICC in this case, since it has a larger penalty function in the range of candidate models (the results are not presented here). In the case where the signal-to-noise ratio is weak and the sample size is small, none of the criteria perform well because detection of the operating model is difficult and underfitting is a serious problem. When the sample size is large, RIC is comparable with BIC, and since both are consistent criteria both perform well (see Section 3). However, of the two, RIC's larger penalty function may resist overfitting too strenuously. As a result it will be more prone to underfitting, and thus its performance is slightly inferior to that of BIC. Next, we give four analytical examples to illustrate how RIC can be used with various model structures.

2.3. *Analytical examples*

In this section we obtain RIC for four models:

(a) a regression model with constant variance;
(b) a Box–Cox transformation model;
(c) a weighted regression model;
(d) a regression model with ARMA errors.

2.3.1. *Example 1 (regression model with constant variance)*

In the case where $W(\theta) = I$ and $\lambda = 1$, equation (2.1) represents a multiple-regression model with constant variance, and the resulting selection criterion is

$$\mathrm{RIC} = (n-k)\log(\tilde{\sigma}^2) + k\{\log(n) - 1\} + \frac{4}{n-k-2}. \tag{2.11}$$

2.3.2. *Example 2 (Box–Cox transformation model)*

Consider $W(\theta) = I$. Then equation (2.1) represents a Box–Cox transformation model with constant variance, and the resulting selection criterion is the same as equation (2.11), except that $\tilde{\sigma}^2$ is evaluated at $\lambda = \hat{\lambda}$ rather than at $\lambda = 1$ as in equation (2.11).

2.3.3. *Example 3 (regression model with non-constant variance)*

Assume that $\lambda = 1$ and $W(\theta) = \mathrm{diag}\{W_i(\theta)\}$, where $W_i(\theta) = W(z_i, \theta)$, z_i is a known $q \times 1$ vector and θ is an unknown $m \times 1$ vector for $i = 1, \ldots, n$. Under this assumption, equation (2.1) is a multiple-regression model with non-constant variance. The resulting selection criterion RIC has the same form as equation (2.10) except that the transformation parameter is known to be 1. Other applicable work in this area includes that of Verbyla (1993), who used the residual log-likelihood to obtain a test statistic for assessing the homogeneity of the errors, and that of Lyon and Tsai (1996), who compared various tests for homogeneity obtained using the log-likelihood and the residual log-likelihood.

2.3.4. *Example 4 (regression model with autoregressive moving average errors)*

Consider the regression model (2.1), assuming that $\lambda = 1$ and the random errors are generated by an ARMA(p, q) process defined as

$$e_t - \phi_1 e_{t-1} - \ldots - \phi_p e_{t-p} = a_t - \varphi_1 a_{t-1} - \ldots - \varphi_q a_{t-q},$$

where a_t is a sequence of independent normal random variables having mean 0 and variance σ^2. Let $\theta = (\phi_1, \ldots, \phi_p, \varphi_1, \ldots, \varphi_q)'$ and W be the variance matrix of e (see Cooper and Thompson (1977) and Harvey and Phillips (1979)). Then the resulting residual log-likelihood has the same form as equation (2.2) except that λ is replaced by 1 (see Tunnicliffe Wilson (1989) and Cheang and Reinsel (2000)). Therefore, the selection criterion RIC is given by equation (2.10) except that λ is replaced by 1. Note that RIC selects regression parameters by assuming that the orders of p and q in the ARMA process are specified. A different approach was taken by Tsay (1984), who proposed a method to identify the order of p and q when the dimension of the regression parameters is known.

By using arguments similar to those in example 4, RIC can also be obtained for growth curve models and for regression models with spatial correlation errors (Sen and Srivastava (1990), pages 138 and 143). In the next section we address the theoretical justification for applying RIC to the task of the selection of variables.

3. Consistency of the residual information criterion

Here we shall first obtain asymptotic results for RIC for the regression model with constant variance and then extend that result to the regression model with general covariance and the Box–Cox transformation model. Finally, we show that RIC is a consistent criterion in the general model setting.

Let $\mathcal{A} = \{\alpha : \alpha \text{ is a non-empty subset of } \{1, \ldots, k\}\}$ and let

$$\mathcal{A}^0 = \{\alpha \in \mathcal{A} : E\{h(y, \lambda_0)\} = X(\alpha)\beta(\alpha)\}$$

be a subset of $\mathcal{A}$, where each α in $\mathcal{A}^0$ is associated with a model that includes the operating model, and $\{1, \ldots, k\}$ associated with the full model is in $\mathcal{A}^0$. In addition, let $\hat{\alpha}$ denote the model selected by RIC such that its value is the smallest of those for all possible candidate models, and let α^0 be the model in $\mathcal{A}^0$ with the smallest dimension. Thus, the operating model $X_0\beta_0$ can be represented as $X(\alpha^0)\,\beta(\alpha^0)$. To prove asymptotic results, we make the following assumptions.

Assumption 1. $E(\varepsilon_1^{4s}) < \infty$ for some positive integer s.

Assumption 2. $0 < \liminf_{n\to\infty} \min_{\alpha\in\mathcal{A}} |X(\alpha)'\,X(\alpha)/n|$ and $\limsup_{n\to\infty} \max_{\alpha\in\mathcal{A}} |X(\alpha)'\,X(\alpha)/n| < \infty$.

Assumption 3. $\liminf_{n\to\infty}(n^{-1}) \inf_{\alpha\in\mathcal{A}-\mathcal{A}^0} \|X_0\beta_0 - H(\alpha)X_0\beta_0\|^2 > 0$, where

$$H(\alpha) = X(\alpha)\{X(\alpha)'\,X(\alpha)\}^{-1}X'(\alpha).$$

We can now present RIC's asymptotic properties for the regression model with constant variance ($\lambda_0 = \lambda = 1$ and $W(\theta_0) = W(\theta) = I$).

Theorem 1. If assumptions 1–3 are satisfied, $\mathcal{A}_0$ is not empty and the ε_i are independent and identically distributed (IID), then RIC is a consistent criterion (i.e. $P(\hat{\alpha} = \alpha^0) \to 1$ as $n \to \infty$).

The proof of theorem 1 is given in Appendix A.1.

Next, we consider the regression model with a general covariance structure ($\lambda_0 = \lambda = 1$ is known). Suppose that there are matrices $Q(\theta_0)$ and $Q(\theta)$ such that $W(\theta_0) = Q(\theta_0)\,Q(\theta_0)'$ and $W(\theta) = Q(\theta)\,Q(\theta)'$. Then the above asymptotic result can be extended, as we see below.

Theorem 2. Assume that both $\hat{\theta} - \theta_0$ and $\tilde{\sigma}^2(\alpha, \hat{\theta}) - \tilde{\sigma}^2(\alpha, \theta_0)$ tend to 0 in probability as $n \to \infty$ for all $\alpha \in \mathcal{A}$. In addition, assume that the elements of $W(\theta)$ are continuous functions of θ, and $W(\theta)$ is positive definite in the neighbourhood of θ_0. If assumptions 1–3 are satisfied when $X(\alpha)$ and ε are replaced by $Q(\theta)^{-1}\,X(\alpha)$ and $\varepsilon_0 = Q(\theta)^{-1}\varepsilon$ respectively, $\mathcal{A}_0$ is not empty and the ε_{i0} are IID, then RIC is a consistent criterion.

The proof of theorem 2 is given in Appendix A.2. In addition to the above extension, we can obtain the asymptotic results for the Box–Cox transformation model ($W(\theta_0) = W(\theta) = I$) by adding the following assumption.

Assumption 4. Assume that $h(t, \lambda)$ is a continuously differentiable function of λ, and for some constant $c > 0$

$$\sup_{|\lambda|\leqslant c} \left\{ n^{-1} \sum_{i=1}^{n} h'(y_i, \lambda)^2 \right\} = O_p(1),$$

where $h'(t, \lambda) = \partial h(t, \lambda)/\partial\lambda$.

This gives us the following result, the proof of which is in Appendix A.3.

Theorem 3. If assumptions 1–4 are satisfied, $\mathcal{A}_0$ is not empty, the ε_i are IID and $\hat{\lambda}$ is a root n consistent estimate of λ_0, then RIC is a consistent criterion.

Theorems 1–3 remain true when BIC is substituted for RIC. By incorporating all the conditions stated in these theorems, we have the asymptotic results of RIC for the general model setting (2.1).

Corollary. Suppose that both $\hat{\theta} - \theta_0$ and $\tilde{\sigma}^2(\alpha, \hat{\theta}) - \tilde{\sigma}^2(\alpha, \theta_0)$ tend to 0 in probability as $n \to \infty$ for all $\alpha \in \mathcal{A}$. Assume that the elements of $W(\theta)$ are continuous functions of θ, and $W(\theta)$ is positive definite in a neighbourhood of θ_0. If $\mathcal{A}_0$ is not empty, $\hat{\lambda}_0$ is a root n consistent estimate of λ_0, assumptions 1–3 are satisfied when $X(\alpha)$ and ε are replaced by $Q(\theta)^{-1}\,X(\alpha)$ and ε_0 respectively, ε_{i0} are IID and assumption 4 holds, then RIC is a consistent criterion.

4. Simulations

In this section we use simulation studies to compare the performance of RIC *versus* AIC $= n\log(\hat\sigma^2)+2k$, $C_p = n\hat\sigma^2/s^2 - n + 2k$,

$$R^2_{\text{adj}} = 1 - \left\{(n-1)\tilde\sigma^2 \Big/ \sum_{i=1}^{n}(y_i-\bar y)^2\right\},$$

FPE $= n\hat\sigma^2(n+k)/(n-k)$, AICC and BIC, where s^2 is the unbiased estimate of σ^2 computed under the full model (including all covariates).

We shall consider four regression models for these simulations—models with constant variance, transformation, non-constant variance and AR(1) errors—which parallel the analytical examples 1–4. For the transformation model, 1000 realizations were generated from the operating model, which can be viewed as the true model for the purpose of simulation studies. In this model, λ_0 ranges from -2 to 2 in increments of 0.5, $\beta_0 = (a_0, 1, 1, 1)'$ and $\sigma_0^2 = \sigma_\mu^2/\gamma_0^2$, where σ_μ^2 is the variance of the mean function of the operating model, $a_0 = 8$ if $\lambda_0 \geqslant 0$ and $a_0 = -8$ if $\lambda_0 < 0$. These values of a_0 were chosen to ensure that the response $h(y, \lambda_0)$ is a vector of real values. The quantity γ_0 is the signal-to-noise ratio. For the remaining three models, 1000 realizations were generated from their corresponding true models with $\beta_0 = (1, 1, 1, 1)'$ and $\sigma_0^2 = \sigma_\mu^2/\gamma_0^2$.

In both the constant variance and the transformation models, the ε_i are IID normal ran-

Table 1. Percentages of correct model order selection by AIC, AICC, BIC, C_p, FPE, RIC and R^2_{adj} in 1000 realizations for the usual regression model with constant variance and signal-to-noise ratios 1, 3 and 5

n	*Criterion*	*% correct selection for the following signal-to-noise ratios:*		
		1	*3*	*5*
20	AIC	21	44	41
	AICC	18	78	75
	BIC	18	63	60
	C_p	20	57	56
	FPE	21	48	44
	RIC	23	86	91
	R^2_{adj}	18	28	25
40	AIC	47	54	53
	AICC	59	69	67
	BIC	61	80	78
	C_p	53	60	58
	FPE	48	55	53
	RIC	60	92	95
	R^2_{adj}	30	31	29
80	AIC	59	54	57
	AICC	65	62	65
	BIC	87	87	88
	C_p	61	57	60
	FPE	59	54	57
	RIC	81	94	97
	R^2_{adj}	33	32	30

dom variables with mean 0 and variance σ_0^2. In the non-constant variance model, the ε_i are independent normal random variables with mean 0 and variance $\sigma_0^2 W_0$, where the ith diagonal element of W_0 is $W(z_i, \theta_0) = \exp(\theta_0 z_i)$, the z_i are IID standard normal and the θ_0-values are 0.2, 0.5 and 0.7. In the regression model with AR(1) errors, the ε_i are random errors satisfying $\varepsilon_1 \sim N\{0, 1/(1 - \theta_0^2)\}$, $\varepsilon_i = \theta_0 \varepsilon_{i-1} + \delta_i$, $\delta_i \sim N(0, 1)$, for $i = 2, 3, \ldots, n$, and $\theta_0 = 0.1$, 0.5, 0.9.

Six candidate variables were stored in an $n \times 6$ matrix $\tilde{X}$ of IID uniform random variables $U(-1, 1)$ and standard normal random variables $N(0, 1)$ respectively for the transformation and non-transformation models. The candidate models are linear and include all subsets of the columns of $X = (\mathbf{1}, \tilde{X})$, except for the empty set. The operating model is described by X_0, the first k_0 $(= 4)$ columns of X. In the simulations that follow, the signal-to-noise ratio values γ_0 are 1, 3 and 5. In addition, three sample sizes, $n = 20, 40, 80$, were used. For the transformation model, we use the maximum likelihood estimate computed from the full model to estimate λ since it is a root n consistent estimate of λ_0 (see theorem 5.3 of Bickel and Doksum (1981)). Analogously, in both the non-constant-variance model and the regression model with AR(1) errors, we use the maximum likelihood estimate computed from the full model to estimate θ because it is a consistent estimate of θ_0.

Tables 1–4 give the percentages of true model selection by the various criteria for each of the regression models (constant variance, transformation, non-constant variance and AR(1)

Table 2. Percentages of correct model order selection by AIC, AICC, BIC, C_p, FPE, RIC and R^2_{adj} in 1000 realizations of Box–Cox transformation models when the transformation parameter of the candidate model is unknown and the signal-to-noise ratios are 1, 3 and 5

n	*Criterion*	*% correct selection for the following signal-to-noise ratios:*		
		1	*3*	*5*
20	AIC	26	42	40
	AICC	25	73	73
	BIC	26	60	58
	C_p	28	55	52
	FPE	27	45	42
	RIC	17	84	92
	R^2_{adj}	21	26	25
40	AIC	47	51	51
	AICC	60	67	66
	BIC	66	78	78
	C_p	52	58	57
	FPE	47	52	51
	RIC	61	95	96
	R^2_{adj}	29	30	29
80	AIC	55	57	54
	AICC	62	63	60
	BIC	86	87	85
	C_p	58	60	56
	FPE	55	57	54
	RIC	85	96	97
	R^2_{adj}	31	31	30

respectively). In Table 2, we present only the case where $\lambda_0 = 0$ since the performance of each criterion is similar across all data sets simulated from various values of the true transformation parameter. Tables 1 and 2 show that RIC outperforms the other criteria when $\gamma_0 = 3$ and $\gamma_0 = 5$. When $\gamma_0 = 1$ and $n = 20$, none of the criteria performs well although RIC performs worse than the others with the Box–Cox transformation model. As the sample size increases to 40 or 80, however, RIC's performance improves with increasing n such that it is comparable with the best selection criterion, BIC.

AICC was proposed by Hurvich and Tsai (1989) for the constant variance model to overcome the overfitting problem of AIC that arises when the sample size is small. Table 1 shows a 66% improvement for AICC over AIC where $n = 20$ and $\gamma_0 = 3$ or $\gamma_0 = 5$. Moreover, RIC makes another 12% improvement. It is also interesting that AICC performs better than BIC when the sample size is small and the signal-to-noise ratio is strong. This is because AICC has a larger penalty function than BIC over the range of candidate models (not presented here). A pattern similar to that described above is also found in the transformation model (see Table 2).

For both non-constant variance and AR(1) models, Tables 3 and 4 show findings that are similar to those presented in Tables 1 and 2. However, one additional point should be noted, which is that RIC performs the best in the small sample case ($n = 20$), without regard to the degree of signal-to-noise ratio or the magnitude of θ_0. RIC retains its lead as

Table 3. Percentages of correct model order selection by AIC, AICC, BIC, C_p, FPE, RIC and R^2_{adj} in 1000 realizations for the weighted regression model with θ_0 =0.2, 0.5, 0.7 and signal-to-noise ratios 1, 3 and 5

n	*Criterion*	*% correct selection for the following values of θ_0 and signal-to-noise ratios:*								
		θ_0 = 0.2			*θ_0 = 0.5*			*θ_0 = 0.7*		
		1	*3*	*5*	*1*	*3*	*5*	*1*	*3*	*5*
20	AIC	18	41	41	33	48	45	43	54	53
	AICC	18	71	73	37	76	73	56	79	78
	BIC	18	58	58	36	64	60	51	66	66
	C_p	19	55	54	37	59	56	50	63	63
	FPE	18	42	43	33	50	47	44	56	54
	RIC	23	83	90	42	89	92	57	91	93
	R^2_{adj}	15	27	28	25	34	32	35	40	40
40	AIC	50	51	50	56	58	57	60	63	62
	AICC	61	65	65	70	71	70	70	71	73
	BIC	65	76	77	79	80	81	79	82	81
	C_p	55	57	56	61	63	62	64	65	64
	FPE	51	52	51	57	58	57	61	63	62
	RIC	65	91	95	79	93	96	82	94	95
	R^2_{adj}	32	32	28	36	38	38	44	47	42
80	AIC	57	56	55	58	60	59	65	66	65
	AICC	64	63	62	63	67	64	70	71	71
	BIC	87	88	87	87	88	88	89	87	86
	C_p	60	59	58	60	63	61	66	68	66
	FPE	58	56	55	58	60	59	65	66	65
	RIC	80	94	96	85	95	98	90	95	97
	R^2_{adj}	33	29	30	34	39	36	46	47	45

Table 4. Percentages of correct model order selection by AIC, AICC, BIC, C_p, FPE, RIC and R^2_{adj} in 1000 realizations for the regression model with AR(1) errors, θ_0 = 0.1, 0.5, 0.9 and signal-to-noise ratios 1, 3 and 5

n	*Criterion*	*% correct selection for the following values of θ_0 and signal-to-noise ratios:*								
		θ_0 = 0.1			*θ_0 = 0.5*			*θ_0 = 0.9*		
		1	*3*	*5*	*1*	*3*	*5*	*1*	*3*	*5*
20	AIC	22	45	48	22	48	45	33	46	46
	AICC	18	77	78	21	80	78	48	77	78
	BIC	20	61	64	23	64	64	41	64	63
	C_p	21	57	61	25	59	58	39	59	58
	FPE	22	47	50	23	49	46	34	48	48
	RIC	24	87	94	27	89	93	49	88	96
	R^2_{adj}	18	28	30	18	30	27	24	33	30
40	AIC	48	56	53	47	54	54	46	53	56
	AICC	58	70	70	55	67	67	56	69	71
	BIC	62	80	81	56	78	79	61	81	81
	C_p	52	61	59	50	60	60	51	58	62
	FPE	48	56	53	48	55	55	46	53	57
	RIC	60	91	95	56	90	95	63	94	97
	R^2_{adj}	30	33	30	29	31	31	29	31	32
80	AIC	59	57	58	56	59	54	53	56	58
	AICC	66	63	64	64	65	61	59	65	65
	BIC	85	87	89	83	88	86	74	89	89
	C_p	62	60	61	59	62	57	56	59	61
	FPE	59	57	58	56	59	54	53	57	58
	RIC	80	94	97	78	95	97	73	95	98
	R^2_{adj}	33	32	31	30	33	30	31	30	31

the sample size increases to 40 or 80, except when $\gamma_0 = 1$, where it is comparable with BIC.

All the above Monte Carlo findings can be explained as follows.

(a) RIC performs well in large samples because it is a consistent criterion.

(b) In both the constant variance and the transformation models, detecting the correct model is very difficult when the signal-to-noise ratio is 1 and the sample size is 20. As a result, underfitting is a serious problem and none of the selection criteria performs well.

(c) For a given signal-to-noise ratio, model identification becomes easier as θ_0 increases in the non-constant-variance and AR(1) models. Thus RIC's large penalty function, which prevents overfitting, enables it to perform well.

(d) When the signal-to-noise ratio increases, the model is more easily identified, which leads to an improved performance by RIC.

In addition to these Monte Carlo studies, we also conducted further simulations using the candidate model's parameter estimates of λ and θ to compute the selection criteria. Since these estimates are not consistent, the resulting criteria (including RIC) do not perform well and are not presented here. On the basis of the results of our simulations, we recommend that RIC be considered routinely for regression model selection, except when the sample size is small and the signal-to-noise ratio is weak.

5. Conclusion

We identify the following four research areas for further study in applying the residual log-likelihood approach. First, adapt the approach of Hurvich *et al.* (1998) and the extension of Simonoff and Tsai (1999) to obtain RIC for nonparametric models, semiparametric models, additive models and non-linear models. Second, obtain RIC for time series models and models for longitudinal data (see Jones (1993), Diggle *et al.* (1994) and Cheang and Reinsel (2000)). Third, derive RIC for nonparametric transform models (see He and Shen (1997)) and transform-both-sides models (see Carroll and Ruppert (1988)). Fourth, extend the application of RIC to mixed models (see Brown and Prescott (1999) and McCulloch and Searle (2000)). We believe that these efforts would lead to better model selection methods for data analysis.

Acknowledgements

We thank the referee and Joint Editor for their valuable comments and insightful suggestions. Chih-Ling Tsai's research was supported in part by National Institutes of Health grant DA-01-0433 and National Science Foundation grant DMS-95-10511.

Appendix A

Lemma 1. If the ε_i are IID and assumptions 1–3 are satisfied, then $\lim_n[(1/n)\varepsilon'\{I - H(\alpha)\}X_0\beta_0] = 0$ almostly surely uniformly for $\alpha \in \mathcal{A} - \mathcal{A}^0$.

Proof. The lemma follows from assumptions 1–3, Chebychev's inequality and the Borel–Cantelli lemma.

Detailed derivations of lemma 1 and the three theorems can be found at

`http://www.blackwellpublishers.co.uk/rss/`

A.1. Proof of theorem 1

Without loss of generality, we use y in place of $h(y, 1) = y - \mathbf{1}$ in the proof of theorem 1 below, where $\mathbf{1} = (1, \ldots, 1)'$.

A.1.1. Step 1

For $\alpha \in \mathcal{A}$ and $\alpha^0 \in \mathcal{A}^0$, the corresponding model selection criteria are

$$\mathrm{RIC}(\alpha) = n\log\{\tilde{\sigma}^2(\alpha)\} + k(\alpha)\{\log(n) - 1\} + \frac{4}{n - k(\alpha) - 2} - k(\alpha)\log\{\tilde{\sigma}^2(\alpha)\} \tag{A.1}$$

and

$$\mathrm{RIC}(\alpha^0) = n\log\{\tilde{\sigma}^2(\alpha^0)\} + k(\alpha^0)\{\log(n) - 1\} + \frac{4}{n - k(\alpha^0) - 2} - k(\alpha^0)\log\{\tilde{\sigma}^2(\alpha^0)\}, \tag{A.2}$$

where

$$\tilde{\sigma}^2(\alpha) = \frac{y'\{I - H(\alpha)\}y}{n - k(\alpha)},$$
$$\tilde{\sigma}^2(\alpha^0) = \frac{y'\{I - H(\alpha^0)\}y}{n - k(\alpha^0)},$$

$H(\alpha)$ is defined in assumption 3 and $H(\alpha^0) = X(\alpha^0)(X(\alpha^0)'\,X(\alpha^0))^{-1}X'(\alpha^0)$. When the ε_i are IID and satisfy assumption 1, lemma 1 together with theorem 2.4.2 of Rao and Kleffe (1988) imply that

$$\liminf_n \{\tilde{\sigma}^2(\alpha) - \tilde{\sigma}^2(\alpha^0)\} \geqslant \liminf_n \left[\frac{\beta_0' X_0' \{I - H(\alpha)\} X_0 \beta_0}{n - k(\alpha^0)}\right] \geqslant c_1^2 > 0$$

almost surely for all $\alpha \in \mathcal{A} - \mathcal{A}^0$, where $c_1 > 0$ is a constant such that

$$\frac{\beta_0' X_0' \{I - H(\alpha)\} X_0 \beta_0}{n - k(\alpha^0)} \geqslant c_1^2$$

when n is sufficiently large. This leads to

$$\liminf_n [\tilde{\sigma}^{-2}(\alpha^0)\{\tilde{\sigma}^2(\alpha) - \tilde{\sigma}^2(\alpha^0)\}] \geqslant \liminf_n \{c_1^2/\tilde{\sigma}^2(\alpha^0)\} = (c_1/\sigma_0)^2,$$

almost surely for all $\alpha \in \mathcal{A} - \mathcal{A}^0$. Hence,

$$\liminf_n |n[\log\{\tilde{\sigma}^2(\alpha)\} - \log\{\tilde{\sigma}^2(\alpha^0)\}]| \geqslant \liminf_n [n\{\log(1 + (c_1/\sigma_0)^2\}] > 0, \qquad (A.3)$$

almost surely for all $\alpha \in \mathcal{A} - \mathcal{A}^0$.

Next, it is also easy to see that

$$\left[\frac{4}{n - k(\alpha) - 2} - k(\alpha)\ \log\{\tilde{\sigma}^2(\alpha)\}\right] \Big/ n \to 0,$$

almost surely for all $\alpha \in \mathcal{A}$. This result, in conjunction with equations (A.1) and (A.2) and inequality (A.3), implies that

$$\liminf_n \{\mathrm{RIC}(\alpha) - \mathrm{RIC}(\alpha^0)\} > 0, \qquad (A.4)$$

almost surely for all $\alpha \in \mathcal{A} - \mathcal{A}^0$.

A.1.2. Step 2

Here we show the asymptotic property $\{\mathrm{RIC}(\alpha) - \mathrm{RIC}(\alpha^0)\}/\{\log(n) - 1\}$ as $\alpha \in \mathcal{A}^0$. From assumption 1 and theorem 2.4.2 of Rao and Kleffe (1988), we have that

$$\frac{\mathrm{RIC}(\alpha) - \mathrm{RIC}(\alpha^0)}{\log(n) - 1} = \frac{n[\log\{\tilde{\sigma}^2(\alpha)\} - \log\{\tilde{\sigma}^2(\alpha^0)\}]}{\log(n) - 1} + k(\alpha) - k(\alpha^0) + o(1) \qquad (A.5)$$

when $\alpha \in \mathcal{A}^0$.

Next, applying property 2.4 of Chatterjee and Hadi (1988), page 16, and theorem 2.4.2 of Rao and Kleffe (1988), we obtain

$$n[\log\{\tilde{\sigma}^2(\alpha)\} - \log\{\tilde{\sigma}^2(\alpha^0)\}] = o_p\{\log(n) - 1\}. \qquad (A.6)$$

Finally, combining equations (A.5) and (A.6), we have that

$$\frac{\mathrm{RIC}(\alpha) - \mathrm{RIC}(\alpha^0)}{\log(n) - 1} = k(\alpha) - k(\alpha^0) + o_p(1) > 0 \qquad (A.7)$$

for all $\alpha \in \mathcal{A}^0$ and $\alpha \neq \alpha^0$.

A.1.3. Step 3

From expressions (A.4) and (A.7), we obtain that, except for an event whose probability tends to 0 with n,

$$\mathrm{RIC}(\alpha) > \mathrm{RIC}(\alpha^0) \qquad (A.8)$$

for all $\alpha \in \mathcal{A}$ and $\alpha \neq \alpha^0$. Since $\hat{\alpha}$ is the model satisfying $\mathrm{RIC}(\hat{\alpha}) = \min_{\alpha \in \mathcal{A}}\{\mathrm{RIC}(\alpha)\}$, we have $\mathrm{RIC}(\hat{\alpha}) \leqslant \mathrm{RIC}(\alpha^0)$. This, together with inequality (A.8), implies that $\hat{\alpha} = \alpha^0$ when n is sufficiently large, except for an event whose probability tends to 0 with n.

A.2. Proof of theorem 2

Under the assumptions that the elements of $W(\theta)$ are continuous and that $\hat{\theta} \rightarrow \theta_0$ in probability, we obtain that $\log\{|W(\hat{\theta})|/|W(\theta_0)|\} = o_p(1)$ for all $\alpha \in \mathcal{A}$. Hence, equation (2.10) can be rewritten as

$$\begin{aligned}\mathrm{RIC}(\alpha, \hat{\theta}) = {} & (n-k)[\log\{\tilde{\sigma}^2(\alpha, \theta_0)\} + o_p(1)] + \log|W_0| + k(\alpha)\{\log(n) - 1\} \\ & + \frac{4}{n - k(\alpha) - 2} + o_p(1).\end{aligned}$$

Also,

$$\begin{aligned}\mathrm{RIC}(\alpha^0, \hat{\theta}) = {} & (n-k)[\log\{\tilde{\sigma}^2(\alpha^0, \theta_0)\} + o_p(1)] + \log|W_0| + k(\alpha^0)\{\log(n) - 1\} \\ & + \frac{4}{n - k(\alpha^0) - 2} + o_p(1).\end{aligned}$$

By using the same techniques as those used for the proof of theorem 1, we can show that

$$\mathrm{RIC}(\alpha, \hat{\theta}) > \mathrm{RIC}(\alpha^0, \hat{\theta})$$

for all $\alpha \in \mathcal{A}$ and $\alpha \neq \alpha^0$, except for an event whose probability tends to 0 with n.

A.3. Proof of theorem 3

In the case of the Box–Cox regression model, the model selection criteria RIC for $\alpha \in \mathcal{A}$ and $\alpha^0 \in \mathcal{A}^0$ are given in equations (A.1) and (A.2) respectively, except for replacing $\tilde{\sigma}_2(\alpha)$ and $\tilde{\sigma}_2(\alpha^0)$ by

$$\tilde{\sigma}^2(\alpha) = \frac{h(y, \hat{\lambda}_0)'\{I - H(\alpha)\}h(y, \hat{\lambda}_0)}{n - k(\alpha)}$$

and

$$\tilde{\sigma}^2(\alpha^0) = \frac{h(y, \hat{\lambda}_0)'\{I - H(\alpha^0)\}h(y, \hat{\lambda}_0)}{n - k(\alpha^0)}.$$

Using the root n consistency of $\hat{\lambda}_0$ and assumption 4, we obtain that

$$\{h(y, \hat{\lambda}) - h(y, \lambda_0)\}'\{h(y, \hat{\lambda}) - h(y, \lambda_0)\} = o_p\{\log(n)\}.$$

Applying arguments that are similar to those used in the proof of theorem 1, we can verify that

$$\liminf_n |n[\log\{\tilde{\sigma}^2(\alpha)\} - \log\{\tilde{\sigma}^2(\alpha^0)\}]| \geqslant \liminf_n (n[\log\{1 + (c_1/\sigma_0)^2\}]) > 0, \tag{A.9}$$

almost surely for all $\alpha \in \mathcal{A} - \mathcal{A}^0$, and

$$\left[\frac{4}{n - k(\alpha) - 2} - k(\alpha)\log\{\tilde{\sigma}^2(\alpha)\}\right] \Big/ n \rightarrow 0, \tag{A.10}$$

almost surely for all $\alpha \in \mathcal{A}$, where c_1 is a positive constant. Equations (A.9) and (A.10) imply that

$$\liminf_n \{\mathrm{RIC}(\alpha) - \mathrm{RIC}(\alpha^0)\} \geqslant \liminf_n (n[\log\{1 + (c_1/\sigma_0)^2\}]) > 0, \tag{A.11}$$

almost surely for all $\alpha \in \mathcal{A} - \mathcal{A}^0$. Furthermore, using arguments that are similar to those used in the proof of inequality (A.7), we can show that

$$\frac{\mathrm{RIC}(\alpha)-\mathrm{RIC}(\alpha^0)}{\log(n)-1}=k(\alpha)-k(\alpha^0)+o_p(1)>0 \tag{A.12}$$

for all $\alpha \in \mathcal{A}^0$ and $\alpha \neq \alpha^0$.

From expressions (A.11) and (A.12), we obtain that, except for an event whose probability tends to 0 with n,

$$\mathrm{RIC}(\alpha)>\mathrm{RIC}(\alpha^0) \tag{A.13}$$

for all $\alpha \in \mathcal{A}$ and $\alpha \neq \alpha^0$. Since $\hat{\alpha}$ is the model satisfying

$$\mathrm{RIC}(\hat{\alpha})=\min_{\alpha\in\mathcal{A}}\{\mathrm{RIC}(\alpha)\},$$

we have

$$\mathrm{RIC}(\hat{\alpha})\leqslant \mathrm{RIC}(\alpha^0).$$

This, together with inequality (A.13), implies that $\hat{\alpha}=\alpha^0$ when n is sufficiently large, except for an event whose probability tends to 0 with n.

References

Akaike, H. (1970) Statistical predictor identification. *Ann. Inst. Statist. Math.*, **22**, 203–217.

——— (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and F. Csáki), pp. 267–281. Budapest: Akadémiai Kiadó.

Bickel, P. J. and Doksum, K. A. (1981) An analysis of transformation revisited. *J. Am. Statist. Ass.*, **76**, 296–311.

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *J. R. Statist. Soc.* B, **26**, 211–252.

Brown, H. and Prescott, R. (1999) *Applied Mixed Models in Medicine.* New York: Wiley.

Burnham, K. P. and Anderson, D. R. (1998) *Model Selection and Inference (a Practical Information-theoretic Approach)*. New York: Springer.

Carroll, R. J. and Ruppert, D. (1988) *Transformation and Weighting in Regression*. New York: Chapman and Hall.

Chatterjee, S. and Hadi, A. S. (1988) *Sensitivity Analysis in Linear Regression*. New York: Wiley.

Cheang, W. K. and Reinsel, G. C. (2000) Bias reduction of autoregressive estimates in time series regression model through restricted maximum likelihood. *J. Am. Statist. Ass.*, **95**, 1173–1184.

Cooper, D. M. and Thompson, R. (1977) A note on the estimation of parameters of the autoregressive-moving average process. *Biometrika*, **64**, 625–628.

Corbeil, R. R. and Searle, S. R. (1976) Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, **18**, 31–38.

Diggle, P. J., Liang, K.-Y. and Zeger, S. C. (1994) *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

Harvey, A. C. and Phillips, G. D. A. (1979) Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika*, **66**, 49–58.

Harville, D. (1974) Bayesian inference for variance components using only error contrasts. *Biometrika*, **61**, 383–385.

He, X. and Shen, L. (1997) Linear regression after spline transformation. *Biometrika*, **84**, 474–481.

Hurvich, C. M., Simonoff, J. S. and Tsai, C.-L. (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Statist. Soc.* B, **60**, 271–293.

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

Jones, R. H. (1993) *Longitudinal Data with Serial Correlation: a State-space Approach*. New York: Chapman and Hall.

Lahiri, K. and Egy, D. (1981) Joint estimation and testing for functional form and heteroscedasticity. *J. Econometr.*, **15**, 299–307.

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

Lyon, J. D. and Tsai, C.-L. (1996) A comparison of tests for heteroscedasticity. *Statistician*, **45**, 337–349.

Mallows, C. L. (1973) Some comments on C_p. *Technometrics*, **15**, 661–675.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. New York: Chapman and Hall.

McCulloch, C. E. and Searle, S. R. (2000) *Generalized, Linear, and Mixed Models*. New York: Wiley.

Patterson, H. D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika*, **8**, 545–554.

Rao, C. R. and Kleffe, J. (1988) *Estimation of Variance Components and Applications*. Amsterdam: North-Holland.

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

Sen, A. and Srivastava, M. (1990) *Regression Analysis—Theory, Methods, and Applications.* New York: Springer.

Simonoff, J. S. and Tsai, C. L. (1999) Semiparametric and additive model selection using an improved Akaike information criterion. *J. Comput. Graph. Statist.*, **58**, 22–40.

Tsay, R. S. (1984) Regression models with time series errors. *J. Am. Statist. Ass.*, **79**, 118–124.

Tunnicliffe Wilson, G. (1989) On the use of marginal likelihood in time series model estimation. *J. R. Statist. Soc.* B, **51**, 15–27.

Verbyla, A. P. (1990) A conditional derivation of residual maximum likelihood. *Aust. J. Statist.*, **32**, 227–230.

——— (1993) Modelling variance heterogeneity: residual maximum likelihood and diagnostics. *J. R. Statist. Soc.* B, **55**, 493–508.