Hacettepe Journal of Mathematics and Statistics
Volume 31 (2002), 89–97

VARIABLE SELECTION WITH AKAIKE INFORMATION CRITERIA : A COMPARATIVE STUDY

Meral Candan Çetin* and Aydin Erar*

Received 20. 06. 2002

Abstract

In this paper, the problem of variable selection in linear regression is considered. This problem involves choosing the most appropriate model from the candidate models. Variable selection criteria based on estimates of the Kullback-Leibler information are most common. Akaike's AIC and bias corrected AIC belong to this group of criteria. The reduction of the bias in estimating the Kullback-Leibler information can lead to better variable selection. In this study we have compared the Akaike Criterion based on Fisher Information and AIC criteria based on Kullback-Leibler.

Key Words: Akaike information criteria, robust selection, Kullback information, variable selection

1. Introduction

The Akaike information criterion and the corrected Akaike information criterion are based on estimators of expected Kullback-Leibler information. For the benifit of the reader we briefly explain Kullback-Leibler and Fisher information.

Kullback-Leibler Information:

Kullback-Leibler (K-L) information is used as a means of discriminating between the true model and the candidate model. Suppose X is a continuous random vector and $f(x/\theta)$ a probability density function of x, where θ is a p-dimensional parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$, $\theta \in \mathbb{R}^p$.

Let θ^* be the true parameter of θ with density function $f(x/\theta^*)$. Kullback-Leibler information, or the generalized entropy B of Boltzmann, measure the closeness of $f(x/\theta^*)$ to $f(x/\theta)$:

$$\begin{aligned} \mathrm{B}(\theta^*;\theta) &= -\mathrm{I}(\theta^*;\theta) \\ &= \mathrm{E}[\log f(x/\theta) - \log f(x/\theta^*)] \end{aligned} \tag{1}$$

*Hacettepe University, Department of Statistics, Beytepe, Ankara, Turkey.

$$= \int f(x/\theta^*) \log f(x/\theta)\, dx - \int f(x/\theta^*) \log f(x/\theta^*)\, dx$$
$$= \mathrm{H}(\theta^*;\theta) - \mathrm{H}(\theta^*;\theta^*), \tag{2}$$

where I denotes K-L information, E is the expected value with respect to the true function $f(x/\theta^*)$ of x and $\mathrm{H}(\theta^*;\theta)$ is the cross-entropy.

K-L information:

$$\mathrm{I}(\theta^*;\theta) = -\mathrm{B}(\theta^*;\theta)$$
$$= \mathrm{H}(\theta^*;\theta^*) - \mathrm{H}(\theta^*;\theta) \tag{3}$$

is minimized instead of maximizing Equation (2). Since $\mathrm{H}(\theta^*;\theta^*)$ is a constant, $\mathrm{B}(\theta^*;\theta)$ is written as follows,

$$\mathrm{B}(\theta^*;\theta) = \int f(x/\theta^*) \log f(x/\theta)\, dx. \tag{4}$$

The derivative of $\mathrm{H}(\theta^*;\theta)$ with respect to θ at $(\theta = \theta^*)$ is equal to Fisher Information [2].

Fisher Information:

Fisher information contributes vastly to statistical estimation and result theory. It is closely related to efficiency and the sufficiency concept. For n independent observations Fisher information is

$$\mathrm{I}_F = n\mathrm{E}_\theta \left\{ \left(\frac{\partial}{\partial\theta} \sum_{i=1}^{n} \ln f(x_i;\theta) \right)^2 \right\}, \tag{5}$$

where $f(x_i;\theta)$ is the density function of the random variable. Fisher information has a non-negative value and measures the amount of information related to θ. Fisher information is directly related to the correction of the unbiased estimation. Fisher information and Kullback's discrimination information have similarities in sufficiency, efficiency, additiveness and grouping of observations [7].

1.1 The Akaike Information Criterion

The Akaike information criterion, AIC, was developed by Akaike [1] to estimate the expected Kullback Leibler information between the model generating the data and a fitted candidate model.

The Akaike information criterion (AIC) is widely used in model selection because it is an estimate of the expected Kullback-Leibler information of a fitted model [6]. AIC is a biased estimator.

AIC is given by

$$\mathrm{AIC} = n(\log \widehat{\sigma}^2 + 1) + 2p, \tag{6}$$

where p and $\widehat{\sigma}^2$ are the number of parameter and the variance of the subsets model, respectively [6]. AIC involves the selection of subsets which minimize Equation (6). AIC is an asymptotically unbiased estimate. AIC is a sample estimate of expected entropy, expected K-L information or expected cross-entropy.

1.2 The Corrected Akaike Information Criterion

A corrected version of AIC is given by Hurvich and Tsai [6] as

$$\text{AICC} = \text{AIC} + \frac{2(p+1)(p+2)}{n-p-2}. \tag{7}$$

AICC, a method originally proposed for linear regression models by Sugiura [10], is asymtotically efficient in both regression and time series. For linear regression, AICC is unbiased, assuming that the candidate family of models includes the true model. AICC has better bias properties than does AIC.

1.3 Robust Akaike Information Criterion

Ronchetti [8] proposed a robust version of AIC. Ronchetti's [9] robust model selection criteria is given as the function,

$$\text{RAIC}(p;\alpha,\rho) = 2\sum \rho(r_{i,p}) + \alpha p, \tag{8}$$

where $r_{i,p} = (y_i - x_i' T_{n,p})/\widehat{\sigma}$ and $\widehat{\sigma}$ is some robust estimate of σ, $T_{n,p}$ is the M-estimator and $\alpha = 2$. We choose ρ as Huber's function,

$$\rho_c(r) = \begin{cases} r^2/2, & \text{if } |r| \le c \\ c|r|^2 - c^2/2, & \text{otherwise,} \end{cases}$$

where c is a function of the contamination ϵ.

$\text{RAIC}(p;\alpha,\rho_c)$, is the generalized Akaike statistics computed under the least favorable errors's distribution with density $g_0(r) = (1-\varepsilon)(2\pi)^{-1/2}e^{\rho_c(r)}$. The extension of AIC to RAIC is the generalization of maximum likelihood estimation to M-estimation [9].

1.4 Akaike Information Criterion with Fisher Information

In this section, we propose a biased correction for AIC used for variable selection. Çetin [3] proposed using Fisher Information as the measure of the discrepancy between the true and approximating models instead of Kullback-Leibler Information.

Fisher information is now,

$$\mathrm{i}_n(\theta) = n\mathrm{E}_\theta\left[\frac{\partial \log g(y;\theta)}{\partial\theta}\right]^2,$$

where $g(y;\theta)$ denotes the likelihood function under the approximating model, and

$$g_{\theta,\sigma^2}(y) = \frac{1}{(2\pi\sigma^2)}\exp\left[-\sum\nolimits_{i=1}^{n}\frac{(y_i - h(\theta))^2}{2\sigma^2}\right].$$

Taking the derivative of the likelihood function with respect to $h(\theta)$, we obtain

$$\mathrm{i}_n(\theta) = \frac{n}{2}\mathrm{E}_F\left[\frac{(\mu+\epsilon-h(\theta))'(\mu+\epsilon-h(\theta))}{\sigma^2}\right]. \tag{9}$$

$\mathrm{E}_F(\mathrm{i}_n(\theta))$, for a real model F, is a reasonable criterion for judging the quality of the approximating family. Of course, $\mathrm{E}_F(\mathrm{i}_n(\theta))$ is unknown, but it can be estimated. We assume that the approximating family includes the true model. This is a strong assumption, but it is also used in the derivation of AIC.

Thus, we can write $\mathrm{E}(\mathrm{i}_n(\theta))$ as follows,

$$\mathrm{E}(\mathrm{i}_n(\theta)) = \frac{n}{\widehat{\sigma}^2}\left\{\frac{np}{n-p-2}+2\right\}. \tag{10}$$

AIC was designed to provide an approximately unbiased estimate of $\mathrm{E}(\mathrm{i}_n(\theta))$. Consequently,

$$\mathrm{AICF} = \frac{n}{\widehat{\sigma}^2}\left\{\frac{np}{n-p-2}+2\right\} \tag{11}$$

is an approximately unbiased estimator of $\mathrm{E}(\mathrm{i}_n(\theta))$.

2 Simulation Study

We now present a simulation study in order to compare the robust and classical Akaike variable selection criteria. A program was coded using S-Plus functions [3] in order to obtain the percentages with which subsets were chosen by the criteria. We generated 15 independent replicates of five independent uniform random variables on $U[-1,+1]$ and 15 independent normally distributed errors e_i with expectation 0 and variance $\sigma^2 = 0.01, 1, 100$, respectively. Then we generated observations y_i according to the 2 models:

$$\text{Beta1} : (2\sqrt{5}, 4, \sqrt{3}, 1, 0) \text{ and Beta2} : (2\sqrt{5}, 4, -\sqrt{3}, 1, 0).$$

The model is designed to be used without intercept. In order to see the effects of the outliers on the estimators and selection criteria, variable selection criteria are examined in the case of no outlier, one outlier and two outliers for ϵ. We obtained $2^5 = 32$ subsets. However, in the following tables, only the most selected subsets out of the 32 are given. Also, the frequencies of the order selection criteria according to σ^2 are given in the following tables.

In these tables, 'P' shows the model. For example, a '1' labels subsets containing the variable x_1.

Table 1. Percentage of subsets selected by the criteria for Beta1

		$\sigma^2 = 0.01$			$\sigma^2 = 1$			$\sigma^2 = 100$		
	P	RAIC	AICF	AICC	RAIC	AICF	AICC	RAIC	AICF	AICC
No Outlier	1	87.2	-	-	41.6	-	-	23	-	78.0
	12	-	-	0.22	0.2	-	-	16	1	21
	13	-	-	-	12.0	-	-	16	-	-
	123	-	-	-	17.8	-	-	7.6	-	0.6
	124	-	-	-	7.6	-	-	9	-	-
	125	-	-	-	0.4	-	-	5.2	-	0.4
	1234	12.8	100	99.7	5.0	100	100	6.6	38.4	-
	1235	-	-	-	14.2	-	-	3.4	26	-
	1245	-	-	-	1.2	-	-	5.4	27.4	-
	1345	-	-	-	-	-	-	7.8	7.2	-
One Outlier	1	96.8	-	-	56.0	-	-	10.2	-	100
	12	-	-	0.4	-	-	0.6	0.8	-	-
	13	-	-	-	12.8	-	-	14.8	-	-
	123	-	-	-	-	-	-	2.8	-	-
	124	-	-	-	-	-	-	2	-	-
	125	-	-	-	-	-	-	1.2	-	-
	1234	3.2	100	99.6	0.6	100	99.4	14.2	47.2	-
	1235	-	-	-	0.6	-	-	7	2	-
	1245	-	-	-	-	-	-	0.2	-	-
	1345	-	-	-	-	-	-	46.8	50.8	-
Two Outliers	1	99.6	-	-	49.6	-	-	21.8	-	98.8
	12	-	-	-	-	-	32.2	7.8	-	-
	13	-	-	-	49.0	-	-	19.8	0.2	8.2
	123	-	-	-	0.4	-	30.6	11	-	-
	124	-	-	-	0.6	-	-	6.2	-	-
	125	-	-	-	-	-	-	4.8	-	-
	1234	0.4	100	100	-	100	37.2	5.8	24.4	-
	1235	-	-	-	-	-	-	6.6	42.8	-
	1245	-	-	-	0.4	-	-	9	2	-
	1345	-	-	-	-	-	-	7.2	30.6	-

Table 2. Percentage of subsets selected by the criteria for Beta2

		$\sigma^2 = 0.01$			$\sigma^2 = 1$			$\sigma^2 = 100$		
	P	RAIC	AICF	AICC	RAIC	AICF	AICC	RAIC	AICF	AICC
No Outlier	1	96.8	-	-	96.8	-	-	28	-	50.4
	12	-	-	-	-	-	-	13.6	-	40
	13	-	-	-	-	-	-	19	-	6.6
	123	-	-	-	-	-	-	8.4	-	0.8
	124	-	-	-	-	-	-	8.6	-	1.4
	125	-	-	-	-	-	-	5.8	-	0.8
	1234	3.2	100	100	3.2	100	100	5.4	39.4	-
	1235	-	-	-	-	-	-	2.4	27.8	-
	1245	-	-	-	-	-	-	3.4	24.8	-
	1345	-	-	-	-	-	-	5.4	8	-
One Outlier	1	99.4	-	-	39.4	-	-	26.2	-	99.8
	12	-	-	-	3.4	-	1.6	3.6	-	0.2
	13	-	-	-	51.8	-	-	8.6	-	-
	123	-	-	-	-	-	0.2	0.4	-	-
	124	-	-	-	0.2	-	63.2	3	-	-
	125	-	-	-	3.8	-	-	4.2	-	-
	1234	0.6	100	100	0.4	100	35.0	10.4	45.6	-
	1235	-	-	-	1.0	-	-	4.2	2	-
	1245	-	-	-	-	-	-	2	4.4	-
	1345	-	-	-	-	-	-	37.4	48	-
Two Outliers	1	100	-	-	31.8	-	-	20.4	1.4	100
	12	-	-	-	3.0	-	32.0	8.2	-	-
	13	-	-	-	56.4	-	-	18.4	0.2	-
	123	-	-	-	-	-	4.2	12.2	-	-
	124	-	-	-	2.8	-	52.8	6.2	-	-
	125	-	-	-	2.8	-	-	6.8	-	-
	1234	-	100	100	-	100	11.0	5	19	-
	1235	-	-	-	-	-	-	11.8	29	-
	1245	-	-	-	3.2	-	-	7.4	31.2	-
	1345	-	-	-	-	-	-	8.6	19.2	-

According to Table 1, when the variance is small, the AICC and AICF criteria select the true model '1234' with 100% without outliers. On the contrary, the RAIC criteria performs poorly in all situations. In the cases of no outlier and one outlier, the results for the AICF and AICC criteria are similar for $\sigma^2 = 0.01$ and $\sigma^2 = 1$. With one outlier, AICF also selects the model '1234' with 100% and AICC selects the model '1234' with 99.4%. When $\sigma^2 = 1$, the AICC criterion fails in the case of two outliers. In general, the results for $\sigma^2 = 0.01$ and $\sigma^2 = 1$ are similar for Beta1. Although the AICC criterion provide good performance for $\sigma^2 = 0.01$ and $\sigma^2 = 1$, it fails for $\sigma^2 = 100$. AICC is known to be affected by a change in variance, and when the variance is large [4]. In the case of large variance, with respect to outliers, the AICF criterion also selects the true model '1234' and the model '1345', whereas AICC fails. When the variance gets bigger, the AICC criterion is affected a lot by outliers. The AICF criterion also selects model '1234' with 100%, even if there are two outliers.

In Table 2, it is seen that the results for Beta2 are similar to those for Beta1. In the case of Beta2, while AICC and AICF provide as good a performance as for Beta1, RAIC again does not provide a good performance. However, in contrast to Beta1, in the case of one outlier the criterion AICC fails for $\sigma^2 = 1$.

An Application using Hald Data

This data set consists of 13 observations on 4 independent variables, where these variables are related to one other. It should be pointed out that there is multicollinearity among the regressors for this data. It has been observed that there are no outliers in the y direction. Some of the results for 15 subsets obtained using Hald Data [5], when there are no outliers, are given in Table 3.

Table 3. Results for some subsets using Hald Data

P	RAIC	AICF	AICC
12	11.49	36.88	36.62
13	150.0	1.74	53.8
14	6.10	28.56	38.07
123	11.22	45.80	42.46
124	9.24	45.93	42.44
134	3.91	43.34	42.77
234	17.28	29.85	44.88

For Hald Data, model '12', that is the model containing x_1 and x_2, is especially preferred as the predicted model. There is not multicollinearity in this subset. Model '124' can be proposed as the best models with 3 variables.

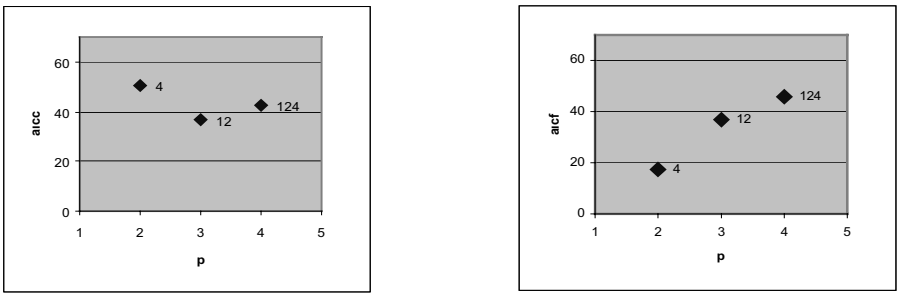
It is seen from Table 3 and Figure 1 that the models '12' and '124' are selected by the AICC criterion. When the model '12' is selected by the AICC criterion, the

'124' model is also a suitable 3 variable model. But, if the AICF value is plotted against the parameter number, it is seen that '12' is selected by the AICF.

AICF selects the model '124' with three variable. It is seen that AICF tends to select subsets with many variables (Çetin, [3]). RAIC also does not provide a good performance in the simulation study.

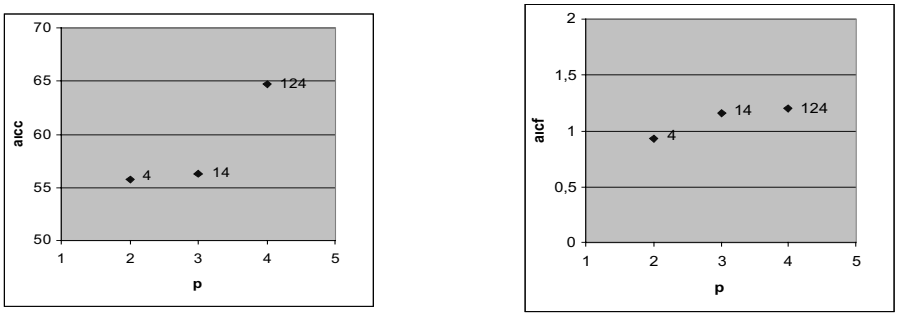
To see the effect of outliers, outliers are generated on the dependent variable: The dependent variable is replaced firstly with $y_6 = 150$, later by $y_6 = 150$ and $y_8 = 150$. The results are given in Figures 2 and 3.

Figure 1. AICC-p and AICF-p without outliers



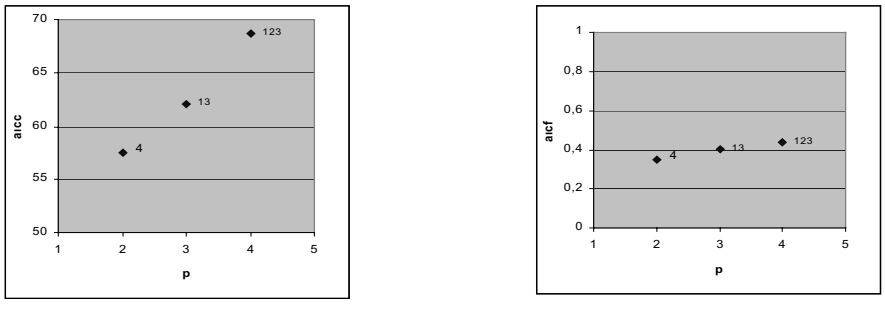
(a) (b)

Figure 2. AICC-p and AICF-p with one outlier



(a) (b)

Figure 3. AICC-p and AICF-p with two outliers



(a) (b)

In the case of one outlier, the AICC and AICF criteria select model '14'. From the former study on Hald Data, it is known that subset '14' is one of the best models. Also, subset '14' is selected by AICC and AICF for one outlier.

According to figure 2, AICC and AICF select the same subset '13'. However, this situation is not much desired for AICC. AICF is affected by two outliers.

References

[1] Akaike, H. A new look at the statistical model identification, IEEE Transaction on Automatic Control **19**, 716–723, 1974.

[2] Bozdoğan, H. Model selection and Akaike's information criterion: The general theory and its analytical extensions, Psychometrika **52 (3)**, 345–370, 1987.

[3] Çetin, M. Sağlam Regresyonda Değişken Seçim Ölçütleri, Unpublished PhD Thesis (Turkish), Hacettepe University, Ankara, 2000.

[4] Çetin, M. and Erar, A. Variable selection criteria in robust regression, ?? **30**, ISSN 1300-4263, 77–83, 2001.

[5] Draper , N. R. and Smith, H. Applied Regression Analysis, Wiley, New-York, 1981.

[6] Hurvich, C. F. and Tsai, C. L. Regression and time series model selection in small samples, Biometrika **76 (2)**, 297–307, 1989.

[7] Kotz, S. and Johnson, N. L. Encylopedia of Statistics Sciences **3**, JohnWiley & Sons, 1982.

[8] Ronchetti, E. Robust Testing in Linear Models: The Infinitesimal Approach, Ph. D. thesis, ETH Zürich, 1982.

[9] Ronchetti, E. Robust model selection in regression, Statistics and Probability Letters. **3**, 21–23, 1985.

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