ELSEVIER Computational Statistics & Data Analysis 22 (1996) 461–469

COMPUTATIONAL STATISTICS & DATA ANALYSIS

Model selection in linear regression

Debasis Kundu*,[1], G. Murali

Department of Mathematics, Indian Institute of Technology, Kanpur, India

Received May 1995; revised October 1995

Abstract

We consider the multiple regression model $\boldsymbol{Y} = \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$ are n-vector random variables, $\boldsymbol{X}_0$ is an $n \times m$ design matrix and $\boldsymbol{\beta}$ is an m-vector of unknown regression parameters. It is well known that different information theoretic criteria with proper choice of penalty function can be used to choose the correct model. In this paper we have done an extensive simulation study to choose the proper penalty function, by using different models and using different error random variables.

Keywords: Model selection; AIC; BIC; EDC; Consistent estimates; Penalized likelihood

1. Introduction

Since the true model is seldom known, the model selection techniques are very useful in linear regression analysis (see for example, Draper and Smith, 1981; Linhart and Zucchini, 1986). We consider the multiple regression model $\boldsymbol{Y} = \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$ are n-vector random variables, $\boldsymbol{X}_0$ is an $n \times m$ design matrix and $\boldsymbol{\beta}$ is an m-vector of unknown regression parameters. We assume that $\boldsymbol{E}(\boldsymbol{\varepsilon}^{\mathrm{T}}) = \mathbf{O}$ and $\boldsymbol{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}) = \sigma^2\boldsymbol{I}_n$. For simplicity, we assume that the candidate models are nested and that the true model is one of the candidate models, i.e. we assume that $\boldsymbol{\beta}$ is of the form: $\boldsymbol{\beta} = \boldsymbol{\beta}(k) = (\beta_1, \ldots, \beta_k \neq 0, 0, \ldots, 0)$.

The most general case would be to consider the multiple regression model $\boldsymbol{Y} = \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where each component of $\boldsymbol{\beta}$ may be zero or non-zero which gives rise to 2^m possible models for multiple regression. But here we are considering the nested hypothesis for simplicity, so we have in total m possible models in our situations.

* Corresponding author. E-mail: kundu@iitk.ernet.in.

[1] Part of the work has been supported by a grant from the Department of Science and Technology, Government of India (No. SR/OY/M-06/93).

0167-9473/96/$15.00 © 1996 Elsevier Science B.V. All rights reserved
PII S0167-9473(96)00008-4

There is a considerable amount of literature on this problem known as selection of variables in regression model; see review papers by Hocking (1976) and Thompson (1978a, b). Recently, methods have been proposed for the choice of a model by minimizing a criterion function defined on the set of alternative models. See for example the works of Akaike (1970, 1973), Mallows (1973), Shibata (1984), Nishi (1984), Schwartz (1978), Risannen (1978), Bai et al. (1986), Rao and Wu (1989) and Kundu (1992, 1995). The main aim of this paper is to choose the proper penalty function by Monte Carlo simulation study using different models and different error random variables.

The organization of the rest of the paper is as follows: in Section 2 we introduce the different information theoretic criteria and in Section 3 we perform the numerical experiments. In Section 4, we draw conclusions from our results.

2. Different criteria

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ be generated from the following process:

$$\boldsymbol{Y} = \boldsymbol{X}_0 \boldsymbol{\beta}(k) + \boldsymbol{\varepsilon}. \tag{2.1}$$

Here $\boldsymbol{X}_0$ is an $n \times m$ known matrix, $\boldsymbol{\beta}(k) = (\beta_1, \ldots, \beta_k \neq 0, \ldots, 0)$ and $\boldsymbol{\beta}(k)$ is an unknown parameter vector, where k is also not known and $\boldsymbol{E}(\boldsymbol{\varepsilon}^{\mathrm{T}}) = \boldsymbol{O}$ and $\boldsymbol{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}) = \sigma^2 \boldsymbol{I}_n$. During the last 10–15 years, several order determination criteria have been proposed to estimate k. Most of the proposed criteria can be expressed in the form:

$$\delta(k) = n \log \hat{\sigma}_k^2 + k g(n), \tag{2.2}$$

where $\hat{\sigma}_k^2$ is the maximum likelihood estimate or the least squares estimate of the residual variance σ^2. In (2.2) the term $kg(n)$ is a non-negative penalty function which increases as the number of parameters increases. On the other hand, $n \log \hat{\sigma}_k^2$ has the tendency to decrease as the number of parameters increases. The value of k that minimizes $\delta(k)$ is the estimate of k. By selecting $g(n) = 2$ in (2.2) we get AIC criterion

$$\mathrm{AIC}(k) = n \log \hat{\sigma}_k^2 + 2k \tag{2.3}$$

introduced by Akaike (1970, 1973). If we select $g(n) = \log n$, the formula (2.2) defines the BIC criterion

$$\mathrm{BIC}(k) = n \log \hat{\sigma}_k^2 + k \log n \tag{2.4}$$

which was defined by Schwartz (1978) and Rissanen (1978). The EDC of Bai et al. (1986) can be expressed as

$$\mathrm{EDC}(k) = n \log \hat{\sigma}_k^2 + k g(n), \tag{2.5}$$

where $g(n)$ satisfies the following properties:

$$\lim_{n \to \infty} \{g(n)/n\} = 0 \quad \text{and} \quad \lim_{n \to \infty} \{g(n)/(\log \log n)\} = \infty. \tag{2.6}$$

It is important to observe that BIC is a special case of EDC. It can be shown Shibata, 1984) that AIC does not give a consistent estimate for the order determination of the model, whereas under the normal error, the BIC and EDC produce a consistent estimate of the order (see Nishi, 1984). The result for the general case is not yet established. It is well known that all information criteria are F-tests with different significance levels (Terasvirta and Mellin, 1986), which depend on the additional power of different columns of the $\boldsymbol{X}$ matrix. It is therefore important to analyze the effect of the correlations of the columns of $\boldsymbol{X}$ on different information criteria.

3. Numerical experiments

We have performed a Monte Carlo simulation study to compare the different criteria, particularly what penalty function should we choose to obtain optimal results. All these computations have been performed on PC-486 using Fortran 77. We consider the following models:

Model 1. $k = 2$, $\boldsymbol{\beta} = (1, 2)$ (lower order)
Model 2. $k = 4$, $\boldsymbol{\beta} = (1, 2, 3, 4)$ (moderate order)
Model 3. $k = 6$, $\boldsymbol{\beta} = (1, 2, 3, 4, 5, 6)$ (large order)

In all these cases there are seven candidate variables for $\boldsymbol{X}$, stored in $n \times 7$ matrix. The $\boldsymbol{X}$ matrix is a fixed matrix. It has been generated once and kept fixed throughout the whole experiment. It has been generated in such a way that each element is normally distributed with mean zero, variance one, the correlation between the two columns of $\boldsymbol{X}$ is ρ and the rows are independent. The candidate models are linear and include the columns of $\boldsymbol{X}$ in a sequentially nested fashion, i.e. the candidate model of dimension k consists of columns $1, \ldots, k$ of $\boldsymbol{X}$. The true design matrix consists of $\boldsymbol{X}_0$, the first k columns of $\boldsymbol{X}$. The error random variables are generated from the four different distributions:

(a) Uniform (bounded and symmetric)
(b) Normal (unbounded and symmetric)
(c) Shifted Beta (bounded and not symmetric)
(d) Shifted Chi-Squared (unbounded and not symmetric)

We have taken the mean to be zeros, different standard deviations, viz., $\sigma = 1$ and 2 and different sample sizes, namely $n = 15$ (small sample), 30 (moderate sample) and 100 (large sample). We have considered different correlation factor of the columns of the matrix $\boldsymbol{X}$, namely $\rho = 0$ (no correlation), $\rho = 0.5$ (moderate correlation), $\rho = 0.9$ (high correlation) and $\rho = 0.95$ (very high correlation). The random deviates are obtained using the algorithm given by Press et al. (1986). Five hundred replications of the data set for different σ and n are generated for all the different models. We mainly compare AIC, BIC and EDC, with different $g(n)$ satisfying (2.6). $g(n) = 2$ gives AIC and $g(n) = \log n$ gives BIC. We take a wide range of $g(n)$ satisfying (2.6) but diverging to infinity at different rate from very slow to very

fast. We take $g(1) = 2$, $g(2) = n^{0.1}$, $g(3) = n^{0.5}$, $g(4) = n^{0.9}$, $g(5) = \log n$, $g(6) = (\log n)^{0.1}$, $g(7) = (\log n)^{0.5}$, $g(8) = (\log n)^{0.9}$, $g(9) = (n \log n)^{0.1}$, $g(10) = (n \log n)^{0.5}$ and $g(11) = (n \log n)^{0.9}$. Similarly as of Kundu (1992). The entry in each table represents the number of times out of five hundred replications different methods pick the model. The columns of the tables correspond to k. We do not report all the results here for scarcity of space. We report the results in Tables 1–4 for

Table 1
Model 2, Normal error, $\rho = 0.0$

	k						
Methods	1	2	3	4	5	6	7
$N = 15$							
$g(1)$	0	0	0	425	42	15	18
$g(2)$	0	0	0	358	60	41	41
$g(3)$	0	0	0	487	12	0	1
$g(4)$	489	0	0	11	0	0	0
$g(5)$	0	0	0	345	65	44	46
$g(6)$	0	0	0	316	64	56	64
$g(7)$	0	0	0	329	64	50	57
$g(8)$	0	0	0	341	65	46	48
$g(9)$	0	0	0	364	59	38	39
$g(10)$	0	0	0	493	7	0	0
$g(11)$	500	0	0	0	0	0	0
$N = 30$							
$g(1)$	0	0	0	467	23	9	1
$g(2)$	0	0	0	432	40	23	5
$g(3)$	0	0	0	500	0	0	0
$g(4)$	496	0	0	4	0	0	0
$g(5)$	0	0	0	440	33	22	5
$g(6)$	0	0	0	381	56	46	17
$g(7)$	0	0	0	408	53	31	8
$g(8)$	0	0	0	435	37	23	5
$g(9)$	0	0	0	437	36	22	5
$g(10)$	0	0	0	500	0	0	0
$g(11)$	500	0	0	0	0	0	0
$N = 100$							
$g(1)$	0	0	0	472	25	2	1
$g(2)$	0	0	0	456	37	5	2
$g(3)$	0	0	0	500	0	0	0
$g(4)$	500	0	0	0	0	0	0
$g(5)$	0	0	0	472	25	2	1
$g(6)$	0	0	0	404	58	30	8
$g(7)$	0	0	0	446	42	8	4
$g(8)$	0	0	0	468	29	2	1
$g(9)$	0	0	0	460	34	4	2
$g(10)$	0	0	0	500	0	0	0
$g(11)$	500	0	0	0	0	0	0

Table 2
Model 2, Normal error, $\rho = 0.5$

	k						
Methods	1	2	3	4	5	6	7
$N = 15$							
$g(1)$	0	0	0	422	45	21	12
$g(2)$	0	0	0	360	62	35	43
$g(3)$	0	0	0	484	15	0	1
$g(4)$	500	0	0	0	0	0	0
$g(5)$	0	0	0	346	63	39	52
$g(6)$	0	0	0	319	63	49	69
$g(7)$	0	0	0	333	62	47	58
$g(8)$	0	0	0	344	63	40	53
$g(9)$	0	0	0	365	62	34	39
$g(10)$	0	0	0	487	13	0	0
$g(11)$	500	0	0	0	0	0	0
$N = 30$							
$g(1)$	0	0	0	473	19	8	0
$g(2)$	0	0	0	434	38	22	6
$g(3)$	0	0	0	499	1	0	0
$g(4)$	500	0	0	0	0	0	0
$g(5)$	0	0	0	440	36	19	5
$g(6)$	0	0	0	384	52	40	24
$g(7)$	0	0	0	407	50	33	10
$g(8)$	0	0	0	435	37	22	6
$g(9)$	0	0	0	439	37	19	5
$g(10)$	0	0	0	500	0	0	0
$g(11)$	500	0	0	0	0	0	0
$N = 100$							
$g(1)$	0	0	0	468	27	4	1
$g(2)$	0	0	0	458	35	6	1
$g(3)$	0	0	0	500	0	0	0
$g(4)$	500	0	0	0	0	0	0
$g(5)$	0	0	0	468	27	4	1
$g(6)$	0	0	0	406	60	23	11
$g(7)$	0	0	0	445	39	11	5
$g(8)$	0	0	0	467	28	4	1
$g(9)$	0	0	0	463	31	5	1
$g(10)$	0	0	0	500	0	0	0
$g(11)$	500	0	0	0	0	0	0

Table 3
Model 2, Normal error, $\rho = 0.90$

	k						
Methods	1	2	3	4	5	6	7
$N = 15$							
$g(1)$	2	0	3	417	45	21	12
$g(2)$	0	0	0	362	58	37	43
$g(3)$	184	0	4	302	9	0	1
$g(4)$	500	0	0	0	0	0	0
$g(5)$	0	0	0	347	61	40	52
$g(6)$	0	0	0	320	62	49	69
$g(7)$	0	0	0	333	63	45	59
$g(8)$	0	0	0	345	62	41	52
$g(9)$	0	0	0	364	59	55	41
$g(10)$	259	0	4	231	6	0	0
$g(11)$	500	0	0	0	0	0	0
$N = 30$							
$g(1)$	0	0	0	473	19	8	0
$g(2)$	1	0	0	436	35	23	6
$g(3)$	92	0	0	407	1	0	0
$g(4)$	500	0	0	0	0	0	0
$g(5)$	0	0	0	444	35	19	5
$g(6)$	0	0	0	384	53	39	24
$g(7)$	1	0	0	408	48	31	13
$g(8)$	1	0	0	437	36	21	6
$g(9)$	1	0	0	439	36	20	5
$g(10)$	278	0	0	222	0	0	0
$g(11)$	500	0	0	0	0	0	0
$N = 100$							
$g(1)$	0	0	0	470	27	2	1
$g(2)$	0	0	0	461	32	6	1
$g(3)$	0	0	0	500	0	0	0
$g(4)$	500	0	0	0	0	0	0
$g(5)$	0	0	0	470	27	2	0
$g(6)$	0	0	0	402	63	21	14
$g(7)$	0	0	0	440	41	12	5
$g(8)$	0	0	0	467	28	4	1
$g(9)$	0	0	0	463	31	5	1
$g(10)$	0	0	0	500	0	0	0
$g(11)$	500	0	0	0	0	0	0

Table 4
Model 2, Normal error, $\rho = 0.95$

	k						
Methods	1	2	3	4	5	6	7
$N = 15$							
$g(1)$	49	0	31	350	39	19	12
$g(2)$	4	0	17	341	59	36	43
$g(3)$	396	0	13	86	5	0	0
$g(4)$	500	0	0	0	0	0	0
$g(5)$	4	0	14	330	60	40	52
$g(6)$	3	0	13	307	61	49	67
$g(7)$	3	0	13	319	63	45	57
$g(8)$	4	0	14	328	61	41	52
$g(9)$	5	0	17	344	59	35	40
$g(10)$	431	0	9	57	3	0	0
$g(11)$	500	0	0	0	0	0	0
$N = 30$							
$g(1)$	2	0	5	466	19	8	0
$g(2)$	1	0	0	435	35	23	6
$g(3)$	417	0	2	80	1	0	0
$g(4)$	500	0	0	0	0	0	0
$g(5)$	1	0	0	440	35	19	5
$g(6)$	0	0	0	384	53	39	24
$g(7)$	1	0	0	407	48	30	14
$g(8)$	1	0	0	436	36	21	6
$g(9)$	1	0	0	438	36	20	5
$g(10)$	485	0	0	15	0	0	0
$g(11)$	500	0	0	0	0	0	0
$N = 100$							
$g(1)$	0	0	0	470	27	2	1
$g(2)$	0	0	0	461	32	6	1
$g(3)$	14	0	0	486	0	0	0
$g(4)$	500	0	0	0	0	0	0
$g(5)$	0	0	0	470	27	2	0
$g(6)$	0	0	0	402	63	21	14
$g(7)$	0	0	0	442	41	12	5
$g(8)$	0	0	0	467	28	4	1
$g(9)$	0	0	0	464	30	5	1
$g(10)$	296	0	0	204	0	0	0
$g(11)$	500	0	0	0	0	0	0

model 2, when the error is normally distributed for $\rho = 0, 0.5, 0.9, 0.95$ and when $\sigma = 1.0$.

4. Conclusions

Comparing the results it is observed that for fixed σ as n increases the performance of all the methods are better for all the models. Although it is expected that as σ decreases the performance should be better, but in our simulation (not reported here) it is not very apparent for most of the methods. In fact, it is observed that the performance of most of the methods does not change with small change in standard deviation, namely, from $\sigma = 1$ to $\sigma = 2$. Comparing the results it is clear that all the penalty functions behave similarly for different models and for different error distributions. It is expected that the proper choice of the penalty function is model independent. Although no theoretical justifications can be given. It seems more work is needed in this area. It is observed that for all the distributions, the performance of the different criteria does not change with the order of the model, i.e. any model, whenever it is true, is just as easy or as difficult to detect as the other two models.. We also observe that it is easier to detect the correct model when the error distribution is bounded or symmetric, particularly when the sample size is small. For large sample size it does not make much difference.

Comparing the effect of correlation of the different columns of $\boldsymbol{X}$, it is observed that when the correlation changes from 0 to 0.5 the effect is not significant but when the correlation is increased to 0.9 or 0.95, the performance of most of the methods drops and the worst affected are $g(3)$ and $g(10)$. Although at moderate or large sample sizes the effect is not felt much for all the methods except $g(3)$ and $g(10)$, whereas the performance of $g(3)$ and $g(10)$ are not very satisfactory when the correlation is very high even for large sample sizes.

Finally, it reduces to choosing the proper $g(n)$ which is one of the most important problems in practice. Calculating $g(n)$ at $n = 15, 30$ and 100 allows the various choices to be ranked in order of increasing severity as follows:

$$g(6), g(7), g(8), g(5), g(2), g(9), g(1), g(3), g(10), g(4), g(11).$$

Out of these $\{g(8), g(5)\}$, $\{g(2), g(9)\}$ and $\{g(3), g(10)\}$ behave very similarly. The simulations show that the least severe penalties (especially $g(6)$, $g(7)$) are too weak, causing k to be generally overestimated. The most severe penalty (especially $g(11)$) on the other hand is too severe and k is then generally underestimated and it is completely out of any serious consideration owing to its all prevailing failure. An optimal choice will lie somewhere between these extremes. The simulations also show that $\{g(2), g(9)\}$ and $g(1)$ are a little too weak and $g(4)$ is too severe particularly when the error is large. We recommend $g(3) = (n^{0.5})$ or $g(10)$ $(= n \log n)^{0.5}$ particularly at low or moderate correlation of the columns of $\boldsymbol{X}$. Due to simplicity anyone of them can be used. For high or very high correlation $g(3)$ behaves better than $g(10)$. It is interesting to observe that at high or very high

correlation the performance of $g(3)$ or $g(10)$ is not very satisfactory, particularly at small sample sizes, whereas Akaike criterion works quite well even when the correlation is very high and the sample size is small. In fact, in our experiment we observe that Akaike criterion behaves quite well even for large sample sizes and for all the situations. So it can also be used particularly when the correlation of the columns of $\boldsymbol{X}$ is high or very high.

Acknowledgements

The authors are thankful to two anonymous referees for their very constructive suggestions.

References

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

Akaike, H., Information theory and an extension of the maximum likelihood principle, *Proc. 2nd Internat. Symp. on Information Theory* (Academia Kiado, Budapest, 1973) 267–281.

Bai, Z.D., P.R. Krishnaiah and L.C. Zhao, On the detection of the number of signals in the presence of white noise. *J. Multivariate Anal.* **20** (1986) 1–25.

Draper, N.R. and H. Smith, *Applied Regression Analysis*, 2nd-ed. (Wiley, New York, 1981).

Hocking, R.R., The analysis and selection of variables in linear regression, *Biometrics* **32** (1976) 1–49.

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

Kundu, D., Detecting the number of signals for undamped exponential signals using information theoretic criteria, *J. Statist Comput. Simulation* **44** (1992) 117–131.

Kundu, D., Small sample properties of some parametric and nonparametric methods for detection of signals by Monte Carlo simulations, *IEE Proc. Vision Image Signal Process.* **142**(3)(1995) 181–186.

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

Nishi, R., Asymptotic properties of criteria for selection of variables in multiple regression, *Ann. Statist.* **12** (1984) 758–765.

Press, W.M., S.A. Teukolsky, W.T. Vellerling and B.P. Flannery, *Numerical Recipes in Fortran*, 2nd edn., (Cambridge University Press, Cambridge, 1986).

Rao, C.R. and Y. Wu, A strongly consistent procedure for model selection in a regression problem, *Biometrika* **72**(2) (1989) 369–374.

Rissanen, J., Modeling by shortest data description, *Automatica* **14** (1978) 465–471.

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

Shibata, R., Approximate efficiency of a selection procedure for the number of regression variables, *Biometrika* **71** (1984) 43–49.

Terasvirta, T. and I. Mellin, Model selection criteria and model selection tests in regression models, *Scand. J. Statist.* **13** (1986) 159–171.

Thompson, M.L., Selection of variables in multiple regression, Part I, *Int. Statist. Rev.* **46** (1978a) 1–19.

Thompson, M.L., Selection of variables in multiple regression, Part II, *Int. Statist. Rev.* **46** (1978b) 126–146.