

이학석사 학위논문

비직교 회귀분석에서의 AIC, SICc, Cp 에
대한 연구

A study of AIC, SICc, Cp in Non-Orthogonal Regression

지도교수 박 찬 근

2004년 2월

한국해양대학교 대학원

응 용 수 학 과

류 승 훈

本 論 文 을 류 승 훈 의 理 學 碩 士 學 位 論 文 으 로 認 准 함 .

위 원 장 이 학 박 사 박 춘 일 (인)

위 원 이 학 박 사 박 찬 근 (인)

위 원 이 학 박 사 김 익 성 (인)

2003년 12월 17일

한국해양대학교 대학원

CONTENTS

1. Introduction	1
2. Survey of literature review	2
3. Simulation Study	4
4. Non-orthogonal case	11
5. Conclusion and Further Research	12
5.1 The correlation $\rho = 0.4$	12
5.2 The correlation $\rho = 0.9$	14
abstract	24
reference	25

A study of AIC , $SICc$ and Cp in Non-Orthogonal Regression

Seung Hoon Ryu

Department of Applied Mathematics

Graduate School

Korea Maritime University

ABSTRACT

In this paper, we compare the power of AIC , $SICc$, and Cp to find the true regression model when the independent variables are non-orthogonal. we find the powers and p -values when the sample size is 10 and 25. Also, we simulate the efficiency of each candidate model. We decide the correlation are 0.4 and 0.9.

1. Introduction

One of the most important goals in regression analysis is to find the best model in terms of selection of independent variables. Also there are so many model selection methods and many statisticians are interested in finding new model selection methods. We will compare the power of some model selection criteria. We discuss on the model selection criteria *AIC* (Akaike's Information Criterion, Akaike, 1973, 1978), *SICC* (corrected version of Schwarz Information Criterion, McQuarrie, 1999), and *C_p* (Mallows, 1973) to the non-orthogonal *X* design. We propose using the distribution in the reduction in *SSE* (Sum of Squared Error) for adding one variable as a means of identifying underfit and overfit models. We begin with a review of the non-orthogonal regression model, then describe our procedure.

First, we define the true model and candidate models.

1. The true model is

$$\begin{aligned} Y &= X_* \beta_* + \sigma_* \\ &= X_0 \beta_0 + X_1 \beta_1 + \varepsilon_* \text{ with } \varepsilon_* \sim \mathcal{N}_n(0, \sigma^2 I_n) \end{aligned}$$

2. The overfit candidate model is

$$\begin{aligned} Y &= X \beta + \varepsilon \\ &= X_0 \beta_0 + X_1 \beta_1 + X_2 \beta_2 + \varepsilon \text{ with } \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n) \text{ and } \sigma^2 = \sigma_*^2. \end{aligned}$$

3. The underfit candidate model is

$$Y = X_0 \beta_0 + \varepsilon \text{ with } \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n) \text{ and } \sigma^2 = \sigma_*^2.$$

$$\text{where } Y = (y_1, \dots, y_n),$$

$$X_* = (X_0 \ X_1),$$

$$\varepsilon_* = (\varepsilon_{*1}, \dots, \varepsilon_{*n}), \text{ and}$$

$$\varepsilon = (\varepsilon_1, \dots, \varepsilon_n).$$

We denote the general candidate model by $Y = X\beta + \varepsilon$. In the underfit model, $\hat{\beta}_0$ is unbiased in the orthogonal case, but in general, the $\hat{\beta}_0$ is not unbiased since $E(\hat{\beta}_0) = \beta_0 + (X_0'X_0)^{-1}X_0'X_1\beta_1$. Also, in general, $S^2 = \frac{SSE}{n-k}$ is not unbiased for σ^2 since $E(S^2) > \sigma^2$. Underfit models tend to be too simplistic and make poor predictions.

The overfit candidate model is $Y = X_0\beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon$, and the model is needlessly complex. When k , the number of parameters including the intercept, is close to n , the number of observations, we can get biased $\hat{\beta}$. The variance of $\hat{\beta}$ in the overfit model tends to be greater than that of $\hat{\beta}$ in the true model. The variance of $\hat{\beta}$ in the overfit model is large due to increase of multicollinearity in non-orthogonal regression.

The controlling of underfitting and overfitting is an important rule for finding the best model in regression. The parameter estimates, $\hat{\beta}$, are biased in the underfit model, and the variance of parameter estimates is needlessly large in the overfit model. Park and Park(2001) checked the overfitting probabilities of several model selection criteria in regression. The compromise between biased parameter estimates and a large variance of parameter estimates is one method for finding the true model.

2. Survey of literature review

We check some model selection criteria in this chapter. *AIC* is designed to be an asymptotically unbiased estimator of the Kullback-Leibler information(Kullback and Leibler, 1951) of a fitted model. Kullback-Leibler discrepancy (*K-L*) is a measure of closeness between two density functions. The equation of *AIC* is $AIC = n \log(2\pi) + n \log(\hat{\sigma}_k^2) + n + 2(k+1)$; the first

and third terms are not important for model selection, so we can ignore them. AIC simplifies $AIC = n \log(\hat{\sigma}_k^2) + 2(k+1)$ Scaling AIC by n , we get

$$AIC = \log(\hat{\sigma}_k^2) + \frac{2(k+1)}{n}$$

The model which minimizes AIC is considered to be closest to the true model. However, AIC tends to overfit in small samples (Hurvich and Tsai, 1989). Hurvich and Tsai (1989) attained the bias-corrected, in terms of selected order, version of AIC . $AICc$ is a better criterion than AIC to find the true model in small samples. However, $AICc$ is asymptotically equivalent to AIC in large samples.

$$AICc = \log(\hat{\sigma}_k^2) + \frac{(n+k)}{(n-k-2)}$$

We now examine the relationship between AIC and $AICc$. The equation of $AICc$ from AIC is

$$AICc = AIC + \frac{2(k+1)(k+2)}{(n-k-2)} + n$$

When k goes to $n-2$, the second term of the above equation goes to a plausibility. $AICc$ is AIC plus an additional penalty term. After the penalty function of AIC is scaled by $\frac{n}{(n-k-2)}$, we can also write $AICc$ as

$$\log(\hat{\sigma}_k^2) + \frac{2(k+1)}{n} \times \frac{n}{(n-k-2)}$$

Mallow's C_p (Mallows 1973) is another efficient model selection criterion. The equation is

$$C_p = \frac{SSE_k}{S_k^2} - n + 2k$$

where $S_k^2 = \frac{SSE_k}{(n-k)}$ is the MSE of the most complete model.

Like AIC and $AICc$, small values of C_p indicate better models. When the

candidate model has a small C_p and a C_p close to k , where k is the number of independent variables including the intercept, we can say that the candidate model is the best model. When $C_p > k$, MSE (Mean Squared Error) is large and will indicate an underfit and biased model.

Another class of criteria is the consistent criteria. If the true model belongs to the set of candidate models and is of finite order, then a model selection criterion that identifies the true model asymptotically with probability one is said to be consistent (Shibata, 1980).

The consistent criterion we consider is $SICc$ (McQuarrie, 1999). $SICc$ is derived by using the relationship between AIC and $AICc$. The penalty function of $SICc$ is scaled by $\frac{n}{(n-k-2)}$.

$$SICc = \log(\hat{\sigma}_{\hat{k}}^2) + \frac{\log(n)k}{(n-k-2)}$$

3. Simulation Study.

Consider the orthogonal regression model. Without loss of generality, we assume $X'X = nI_n$. We saw that the best one-variable model includes X_j with the largest $\frac{1}{n}(X'_j Y)^2$ and the best two-variable model includes the largest two $\frac{1}{n}(X'_j Y)^2$. We also know that each of the $\frac{1}{n}(X'_j Y)^2$ are independent χ_1^2 (possibly non-central) random variables. Reduction in SSE has a distribution based on the order statistics of independent χ_1^2 random variables. However, the χ_1^2 may have a central or non-central distribution. In the simplest case, we have independent identically distributed χ_1^2 order statistics. Typically, some of the X_j are important (yielding non-central

χ_1^2), and we have independent but not identically distributed χ_1^2 .

If there are several important variables in the multiple regression model, we need to check the maximum of chi-square distribution order statistics. If these chi-square distributions are all central, it is straightforward to derive the distribution for the drop in *SSE*. However, in the mixed case, no closed form solution exists. This distribution can be approximated using Monte-Carlo or parametric bootstrapping.

SSE follows the chi-square distribution. We need to know the distribution of order statistics of the chi-square distribution to check the drop of *SSE*. Suppose there are three central chi-square distributions each with 1 degree of freedom. X_1 , X_2 and X_3 are independent, identically distributed χ_1^2 random variables with a cumulative distribution function $F(x)$ of X_1 . The mean of the maximum among these central chi-square is the following:

$$\begin{aligned} \mu_{\max} &= \int_{-\infty}^{\infty} x f(x) dx \\ &= \int_0^{\infty} x \left(\int_0^x f(t) dt \right)^2 f(x) dx \\ &= \int_0^{\infty} x \left(3 \int_0^x \frac{1}{\Gamma(1/2)\sqrt{2}} e^{-t/2} dt \right)^2 \frac{1}{\Gamma(1/2)\sqrt{2}} \frac{e^{-x/2}}{\sqrt{x}} dx \end{aligned}$$

The above equation has no closed form but can be evaluated numerically. μ_{\max} shows the expected drop in *SSE* if none of the variables are important.

Suppose there are three chi-square distributions of which one is non-central, $\chi_1^2(\lambda)$, and two are central, χ_1^2 . Let $A \sim \chi_1^2(\lambda)$, $B \sim \chi_1^2$, and $C \sim \chi_1^2$. Then, we should observe $A > B > C$ or $A > C > B$ since A is stochastically larger than B and C . The c.d.f(cumulative distribution function) of $\chi_1^2(\lambda)$ has closed form:

$$F(y) = \int_0^y \sum_{k=0}^{\infty} \frac{e^{-\lambda/2} (\lambda/2)^k}{k!} \frac{1}{2\Gamma(\frac{1}{2} + k)} \left(\frac{y}{2}\right)^{k - (\frac{1}{2})} e^{-y/2} dy$$

However, if the non-centrality parameter, λ , is large, the difference between the non-central chi-square distribution and the central chi-square distribution should be large because the mean of $\chi_1^2(\lambda)$ is $1+\lambda$. Even for two independent χ^2 , a $\chi_1^2(\lambda)$ and χ_1^2 , no closed form exists for computing distributions of the order statistics. We use bootstrapping and Monte-Carlo methods to estimate these distributions. If the non-centrality λ is large in non-central $\chi_1^2(\lambda)$, the remaining other central χ^2 distribution behaves as order statistics from sample size $k-1$. This fact is important in estimating the drop in SSE distribution. The theorem below focuses on asymptotic results.

Theorem. Suppose we have a finite true model of order k^* . As $n \rightarrow \infty$, the p-value for true variables (order $k < k^*$) goes to 0.

Proof) Begin the case of $k^* = 1$. There are 1 non-central $\chi_1^2(\lambda)$ and m independent χ_1^2 distributions. These distributions are independent each other.

$$\begin{aligned} &P(\chi_1^2(\lambda) > \text{maximum of } m \chi_1^2) \\ &= P(\chi_1^2 \lambda - \text{maximum of } m \chi_1^2 > 0) \end{aligned}$$

Suppose that X is $\chi_1^2(\lambda)$ and Y is the maximum of $m \chi_1^2$.

$$= P(X > Y)$$

$$\begin{aligned}
&= \int_0^\infty \int_y^\infty f_X(x) f_Y(y) dy dx \\
&= \int_0^\infty \int_y^\infty \frac{e^{-\lambda/2} (\frac{\lambda}{2})^k}{k(k-1)!} \frac{1}{2\Gamma(\frac{m}{2} + k)} \binom{x}{2}^{(\frac{m}{2}) + k - 1} e^{-\frac{x}{2}} \\
&\quad \frac{m}{(\sqrt{2\pi})^m \sqrt{y} e^y} \left[\int_0^y y t^{-1/2} e^{-t/2} dt \right]^{m-1} dx dy
\end{aligned}$$

There is no closed form of the above equation.

We know that $\lambda = \frac{m\beta_j^2}{\sigma^2}$ in orthogonal case where $X'X = nI$. When n goes to ∞ and $\beta_j^2 > 0$, λ goes to ∞ .

Then, we can say that $E[x_1^2(\lambda)]$ goes to 1.

In general, $P(x_1^2(\lambda_1) > x_1^2) \rightarrow 1$ as $n \rightarrow \infty$

The above theorem tells us that, in large samples, the important variables with $\beta_j \neq 0$ will be added to the model first. The x^2 order statistics should be sorted according to their λ_j . Since λ_j is a function of β_j , we need an unbiased estimate of λ_j using $\hat{\beta}_j$.

We now address how to simulate the distribution of the drop in *SSE*. There are two options to simulate. The first option is to generate a model with no important variables. This method is fast but does not reflect the sorting of central and non-central x^2 distributions. The second option is to fit a model, keep the selected (important) variables, and assume that all other variables are not important. Then, we generate data y^* using $y^* = \hat{\beta}_x + \varepsilon^*$ where ε^* follows $\mathcal{N}(0, S^2)$. As shown below, $\hat{\lambda}$ is close to λ when λ is large. This method is slower but better estimates central and non-central order statistics.

If there are non-central x^2 distributions in order statistics distribution of

variables, the difference between central and non-central χ^2 random variables can be large. We need to estimate the non-centrality, λ_j , of a non-central χ^2 distribution from an estimated $\hat{\beta}_j$ to get the probability of difference. We need bootstrapping of the data because there are unknown parameters in non-central chi-square distributions. The non-centrality is dependent on an unknown $\hat{\beta}_j$, so we need to know the expected value of $\hat{\lambda}_j$,

$$\begin{aligned}
 E[\hat{\lambda}_j] &= E\left[\frac{n \hat{\beta}_j^2}{\sigma^2}\right] \\
 &= E\left[\frac{n(X'X)^{-1}X'Y_j((X'X)^{-1}X'Y_j)'}{\sigma^2}\right] \\
 &= E\left[\frac{X'Y_jY_j'X}{n\sigma^2}\right] \\
 &= \frac{1}{n\sigma^2}X' E[Y_jY_j']X_j \\
 &= \frac{1}{n\sigma^2}X'[Var(Y_j) + E(Y_j)E(Y_j)']X_j \\
 &= \frac{1}{n\sigma^2}X'Var(Y_j)X_j + \frac{1}{n\sigma^2}X'X\beta_j\beta_j'X'X_j \\
 &= \frac{1}{n\sigma^2}X'\sigma^2X_j + \frac{1}{n\sigma^2}\beta_j\beta_j' \\
 &= 1 + \frac{n\beta_j^2}{\sigma^2} \\
 &= 1 + \lambda_j.
 \end{aligned}$$

The estimated non-centrality, $\hat{\lambda}_j$, is not an unbiased estimator, but

$$E[\hat{\lambda}_j - 1] = \lambda_j$$

If all SSE follow the central chi-square distribution, then the order

statistics follow the independent and identical case. Asymptotically, if there are some non-central chi-square distributions, order statistics are independent but not identically distributed. The equation of order statistics is messy, and no closed form solution exists.

We propose using parametric bootstrapping to estimate these distributions using the second option discussed. Using the y^* data from the best order K model, we compute the drop in SSE between the k and $k+1$ variable models. We apply the F^* distribution discussed where $F^* = \frac{SSE_y - SSE_{f^*}}{SSE_{f^*}}$. $F^*_{y^*}$ is compared to $F^*_{y^*}$ where $F^*_{y^*}$ is new data and $F^*_{y^*}$ is from original data. We count the number of times $F^*_{y^*} > F^*_{y^*}$. We repeat the simulation M times. The estimated p -value is the count divided by $M(\frac{count}{M})$. When the p -value is close to 0, the variable is important.

Now, we explain the summary of our procedure.

1. Find the best order $k=0,1,\dots,K$ models ; compute, $\hat{\beta}_k$, S^2_k and SSE_k for each of these models ; and identify the set of important variables. Compute

$$F^*_k = \frac{SSE_{k-1} - SSE_k}{SSE_k}. \text{ Let } k=1.$$

2. Generate data $y^* = \hat{\beta}_{k-1} + \varepsilon^*$ where $\varepsilon^* \sim \mathcal{M}(0, S^2_{k-1})$. Find the best $k-1$ and order k models using y^* . Compute $F^*_{y^*k} = \frac{SSE^*_{k-1} - SSE^*_k}{SSE^*_k}$

If $F^*_{y^*k} > F^*_k$ add 1 to C where C is the count.

3. Repeat the second step M times.
4. Estimated p -value for favoring the order k model over the order $k-1$ variable model is C/M . Favor order k the p -value is small.
5. Repeat steps 2-4 for $k=2, \dots, K$.

We generated data using IMSL and FORTRAN and estimated p -value using the above procedure. AIC , $SICc$, and C_p all select a model. The Fortran subroutines and functions are useful for this simulation in STAT IMSL/LIBRARY(1982). We used the FORTRAN program for all the STAT IMSL/LIBRARY. The program incorporates the IMSL subroutines to generate the random numbers as RNNOF, find the best model as RBEST, and find the covariance COVRVC. IN each case, the number of important variables is 0, 4, and 8. We simulated 1000 times to find these tables using FORTRAN IMSL/LIBRARY. We need to define some notations in tables. p^* is the number of important variables ; p is the number of variables not including the intercept ; and $k = Rank(X) = p + 1$.

Results are summarized in tables that consist of power ; p -value; model selection criteria that are AIC , $SICc$, and C_p ; and efficiency of candidate models. The bootstrapping begins with intercept only. We can find the best one-variable model and then generate regression model j^* to find the best two-variable model.

These steps were repeated to find the best variable models and compute the drop in SSE . We found the distribution of the drop of SSE and then found p -values of important variables in Tables 1 to 4. We also computed the powers of finding important variables after looping the simulation. Our procedure was repeated R times to estimate the power.

We found the powers and p -value when the sample size n is 10 and 25. There are 8 variables not including intercept represented as p in the tables at each case.

4. Non-orthogonal case

In this chapter, we consider the model selection criteria AIC , $SICc$, and Cp in non-orthogonal X design. We find the powers and p -values when the sample size is 10, and 25. The alpha values are 0.1, 0.5, and 0.01. Also, we simulate the AIC , $SICc$, Cp and efficiency of each candidate model.

The problem inverting $X'X$ due to related or correlated columns is called multicollinearity. If there is a relationship among the independent variables, there might be a multicollinearity problem in our regression model. Our problem with multicollinearity is that it reduces the contribution of each X_j to Y . Thus, it is more difficult to find the best model using the model selection method. In this chapter we consider an example of the non-orthogonal case. The example is the correlation between X_j and X_{j+1} that is 0.4. and the other example is when the correlation is 0.9 In our simulation study, $X_0=1$ for the intercept and $X_1 \sim \mathcal{M}(0, 1)$. The correlation between X_1 and X_2 is ρ . Now we get that the conditional distribution of $X_2 | X_1 = x_1 \sim \mathcal{M}(\rho x_1, 1 - \rho^2)$. In general, $X_j | X_{j-1} = x_{j-1} \sim \mathcal{M}(\rho x_{j-1}, 1 - \rho^2)$. The correlation matrix of X is

$$\text{Correlation Matrix} = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 & \rho^6 & \rho^7 \\ \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 & \rho^6 \\ \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 \\ \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^6 & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^7 & \rho^6 & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

and X_1, X_2, \dots, X_{k-1} has a multivariate normal distribution.

The variance inflation factor (VIF) measures the impact of

multicollinearity on the variance of β estimates and measures the relationship

of X_j with the remaining columns of X . If the VIF is less than 5, multicollinearity is not a serious problems. The higher the multiple correlation in the regression model, the lower the precision in the estimate of the parameters. We consider the example, $\rho = 0.4$ and $\rho = 0.9$ VIF for this example is given below.

$$VIF \text{ for } 0.4 \text{ correlation} = \begin{pmatrix} 1.19 \\ 1.38 \\ 1.38 \\ 1.38 \\ 1.38 \\ 1.38 \\ 1.38 \\ 1.38 \\ 1.19 \end{pmatrix} \quad VIF \text{ for } 0.9 \text{ correlation} = \begin{pmatrix} 5.26 \\ 9.53 \\ 9.53 \\ 9.53 \\ 9.53 \\ 9.53 \\ 9.53 \\ 9.53 \\ 5.26 \end{pmatrix}$$

If the X_j are not orthogonal, then the difference between the full and reduced SSE may not be a χ^2 distribution. The order statistics are not from an independent sample. Suppose that the best three-variable model is X_1, X_2 , and X_3 . The best four-variable model may include X_1, X_4, X_5 and X_6 . These models are not nested, and the usual partial F-test does not have an F distribution. Therefore, the best j variable model may not be nested with the best $j+1$ variable model. No closed form solutions exist for working with the drops in SSE .

5. Conclusion and Further Research.

5.1 The correlation $\rho = 0.4$

We generate data and present simulation results that consist of powers ; p -values ; model selection criteria of AIC , $SICc$, Cp ; and efficiency of candidate models when the correlation $\rho = 0.4$ in Tables 1 to 4. In our tables, p^* represents the number of important variables, and p represents the number of variables, not including the intercept of the candidate model. We make tables when sample sizes are 10 and 25. The error variance is 0.1. If p -values are small, the variables tend to be important. When model selection criteria have small values, the candidate model is preferred by the model selection criterion. If the efficiency of the candidate model is close to 1, it means that the model is close to the true model. For $n=10$, S^2_K has 1 degree of freedom which can have an inflationary impact on Cp as seen in Tables.

In Table 1, when the number of important variables is 0, all powers are small and all p -values are greater than 0.5. It means there are no important variables in the regression model. The smallest value of AIC is -3.868 when the number of important variables is 0. It looks like AIC overfits in this case. When p is 1, $SICc$ is -2.353 which is the smallest value. $SICc$ is a better method to find the true model than AIC in this case. The smallest value of Cp is 5.63 when the number of important variables is 0.

When the number of important variable is 8 in Table 1, the power are decreasing when p is increasing, and the smallest p -value is 0.047 when p is 1 because the variables are related each other. The smallest value of AIC is -3.294 when p is 8. It means that AIC chose the true model. $SICc$ underfits in this case because the smallest value of $SICc$ is 0.680. Cp looks to choose the right model. In comparing between orthogonal and $\rho = 0.4$, finding the important variables is difficult using powers and p -value in

Table 1.

When p^* is 8 and p is 1 in Table 2, the power are close to 1. The p -value makes a big jump between when p is 1 and when p is 2. It looks a underfitted using powers and p -values AIC strongly overfits when p^* is 0 and 4, and Cp also overfits in this case. $SICc$ finds the correct important variable.

In Table 3, the p -values of the candidate model are close to 0 when p the is p^* However, p -values are zero or almost zero when p is 1 since the variable are related with each other. In table 4, model selection criteria tends to find the true model since the sample size is a little big.

5.2 The correlation $\rho = 0.9$

In this section, we simulate the powers, p -value, AIC , $SICc$, Cp , and efficiency of the candidate model when the correlation coefficient is 0.9. Tables 5 to 6 represent the simulation results of powers, p -value, efficiency, and model selection criteria. We simulated the same method that we did in Section 5.2 in the correlation $\rho = 0.4$ case.

When p^* , the number of important variables, is 0 in Table 5, powers are small, and p -values are high. The smallest values of AIC and Cp are -4.473 and 9.00, respectively. These model selection criteria strongly overfit. The smallest value of $SICc$ is -23.231 when p , the number of variables not including intercept, is 2. $SICc$ overfits, too, but not as seriously as AIC did. Efficiency is 1 when p is 0 in this case.

When p^* is 8 in Table 5, the model underfits if we use p -values. Also $SICc$ underfits in this case. Although the true model has $p^* = 8$ variables,

efficiency is the highest(0.858)for $p=4$.

In Table 6, when p^* is 2, AIC and C_p overfit while $SICc$ underfits a little. Behavior of AIC , $SICc$, and C_p is the same as when the correlation is 0.4 in this case. In Table 7, when p^* is 8, powers are 1, and p -value is 0 when k is 1 because the correlation is high among variables. $SICc$ underfits, and AIC and C_p have the smallest values when p is 8.

The restriction of an area of the real variable is not good using the C_p and AIC . When the candidate model of AIC and C_p is smaller, the candidate model is better. AIC and C_p are decreased, however, when the variable are bigger. Also, C_p is so big when the variable are small that it is not good idea to use the restriction of variables to find the best model. We can restrict the area of a real variable using the p -value and $SICc$.

The efficiency of the real variable is 1 or close to 1, so we can restrict the area or find the real variable using the efficiency. The variable are related with 0.4 correlation, so finding the model is more difficult than the orthogonal case. The efficiency of the candidate model is 1 or close to 1 when it is a real model. Through Table 1 to 4, AIC tend to overfit, and $SICc$ tend to underfit. Also, C_p has a smaller number when the candidate model has more variable.

In the non-orthogonal case, finding the best model using the these tables is a little difficult because there is multicollinearity in the regression model. When the sample size is large and the model is strong, the powers, p -values, and model selection criteria tend to find the correct important variable.

In this paper, we compare the model selection criteria when the sample

size is 10 and 25. We will study this comparison when the sample size is larger than 25. Also, if the ρ is greater than 0.4, we can calculate the powers of model selection criteria. We already recommended the restriction method of candidate models in orthogonal case (Park, 2000). It is somewhat a good idea to use the restriction method of candidate models in non-orthogonal case.

Table 1. Model selection methods when $\beta_j=1$, $\sigma^2=0.1$, $n=10$, and $\rho=0.4$.

p*	p	powers			P-value	AIC	SIC	Cp	EFF
		$\alpha=0.10$	$\alpha=0.05$	$\alpha=0.01$					
0	0	-2.292	-2.236	232.57	1.000
	1	0.200	0.100	0.000	0.524	-2.529	-2.353	110.72	0.198
	2	0.000	0.000	0.000	0.781	-2.657	-2.270	70.32	0.132
	3	0.100	0.100	0.000	0.799	-2.765	-2.030	56.20	0.106
	4	0.100	0.100	0.000	0.671	-3.047	-1.744	23.62	0.097
	5	0.000	0.000	0.000	0.708	-3.227	-0.973	15.96	0.093
	6	0.100	0.100	0.100	0.586	-3.836	0.137	5.63	0.089
	7	0.000	0.000	0.000	0.821	-3.868	3.742	7.20	0.088
	8	0.000	0.000	0.000	0.751	-3.791	15.133	9.00	0.088
2	0	0.991	1.047	214936.12	0.012
	1	0.800	0.700	0.600	0.040	-0.589	-0.413	11854.24	0.075
	2	0.800	0.700	0.400	0.141	-2.309	-1.922	2287.99	1.000
	3	0.100	0.000	0.000	0.623	-2.694	-1.959	943.61	0.394
	4	0.200	0.000	0.000	0.660	-3.139	-1.836	400.92	0.320
	5	0.000	0.000	0.000	0.510	-3.636	-1.382	150.27	0.294
	6	0.000	0.000	0.000	0.593	-4.198	-0.225	53.87	0.285
	7	0.000	0.000	0.000	0.553	-4.830	2.781	15.99	0.279
	8	0.000	0.000	0.000	0.718	-5.539	13.384	9.00	0.280
4	0	1.993	2.049	15214.8	0.013
	1	0.700	0.600	0.500	0.089	0.937	1.113	4566.6	0.043
	2	0.500	0.500	0.200	0.345	-0.061	0.326	820.33	0.175
	3	0.300	0.300	0.200	0.322	-0.862	-0.127	516.73	0.255
	4	0.900	0.700	0.400	0.127	-2.276	-0.973	68.48	0.991
	5	0.000	0.000	0.000	0.775	-2.552	-0.293	31.10	0.783
	6	0.000	0.000	0.000	0.743	-2.773	1.200	27.57	0.711
	7	0.000	0.000	0.000	0.747	-2.929	4.681	15.72	0.688
	8	0.100	0.100	0.000	0.554	-3.480	15.443	9.00	0.641
6	0	2.244	2.300	248090.33	0.013
	1	0.700	0.600	0.400	0.064	1.237	1.412	54505.24	0.044
	2	0.400	0.400	0.300	0.267	0.134	0.521	19127.29	0.151
	3	0.100	0.000	0.000	0.527	-0.359	0.376	4315.95	0.296
	4	0.100	0.100	0.000	0.605	-0.901	0.401	1727.38	0.503
	5	0.000	0.000	0.000	0.477	-1.470	0.783	834.67	0.638
	6	0.600	0.400	0.100	0.319	-2.933	1.040	83.11	0.916
	7	0.100	0.100	0.000	0.597	-3.672	3.938	27.79	0.903
	8	0.000	0.000	0.000	0.623	-4.400	14.523	9.00	0.870
8	0	2.701	2.757	6651.3	0.007
	1	0.900	0.700	0.200	0.047	1.837	2.013	2311.9	0.021
	2	0.200	0.200	0.000	0.518	1.255	1.641	1371.9	0.045
	3	0.200	0.200	0.100	0.354	0.315	1.050	205.09	0.118
	4	0.400	0.300	0.100	0.381	-0.622	0.680	53.68	0.285
	5	0.000	0.000	0.000	0.560	-1.219	1.035	26.31	0.420
	6	0.100	0.000	0.000	0.591	-1.856	2.117	16.18	0.647
	7	0.000	0.000	0.000	0.557	-2.489	5.121	10.27	0.774
	8	0.100	0.000	0.000	0.484	-3.294	15.629	9.00	1.000

p*: number of important variables. p: number of variables not including intercept

Note that $k = \text{Rank}(X) = p + 1$

Table 2. Model selection methods when $\beta_j=1$, $\sigma^2=0.1$, $n=25$, and $\rho=0.4$.

p*	p	powers			P-value	AIC	SIC	Cp	EFF
		$\alpha=0.10$	$\alpha=0.05$	$\alpha=0.01$					
0	0	-2.288	-2.234	-0.47	1.000
	1	0.000	0.000	0.000	0.663	-2.305	-2.185	-0.61	0.171
	2	0.000	0.000	0.000	1.000	-2.301	-2.102	-0.14	0.116
	3	0.000	0.000	0.000	0.883	-2.287	-1.994	0.62	0.094
	4	0.000	0.000	0.000	0.870	-2.252	-1.847	1.82	0.084
	5	0.000	0.000	0.000	0.923	-2.189	-1.652	3.54	0.081
	6	0.000	0.000	0.000	0.717	-2.130	-1.438	5.18	0.078
	7	0.000	0.000	0.000	0.734	-2.057	-1.182	7.06	0.077
	8	0.100	0.100	0.000	0.668	-1.981	-0.890	9.00	0.075
2	0	1.040	1.094	751.49	0.004
	1	1.000	1.000	1.000	0.000	-0.275	-0.155	184.72	0.014
	2	1.000	1.000	1.000	0.000	-2.222	-2.023	4.38	1.000
	3	0.000	0.000	0.000	0.872	-2.276	-1.983	3.25	0.462
	4	0.000	0.000	0.000	0.833	-2.307	-1.903	3.00	0.351
	5	0.000	0.000	0.000	0.838	-2.277	-1.741	4.10	0.309
	6	0.000	0.000	0.000	0.750	-2.234	-1.542	5.47	0.292
	7	0.000	0.000	0.000	0.686	-2.176	-1.301	7.11	0.284
	8	0.000	0.000	0.000	0.493	-2.102	-1.012	9.00	0.282
4	0	1.681	1.735	1421.9	0.004
	1	1.000	1.000	0.900	0.005	0.766	0.886	486.97	0.012
	2	0.500	0.400	0.200	0.196	0.090	0.289	218.75	0.026
	3	0.300	0.300	0.200	0.265	-0.534	-0.241	100.13	0.054
	4	1.000	1.000	1.000	0.000	-2.125	-1.721	5.74	1.000
	5	0.000	0.000	0.000	0.614	-2.217	-1.681	4.35	0.605
	6	0.000	0.000	0.000	0.794	-2.193	-1.501	5.40	0.540
	7	0.000	0.000	0.000	0.776	-2.134	-1.259	7.05	0.520
	8	0.000	0.000	0.000	0.685	-2.057	-0.966	9.00	0.517
6	0	2.251	2.305	2797.02	0.004
	1	1.000	1.000	1.000	0.000	1.503	1.623	1144.07	0.010
	2	0.100	0.000	0.000	0.484	0.985	1.183	626.02	0.018
	3	0.300	0.200	0.100	0.381	0.532	0.825	361.56	0.030
	4	0.400	0.400	0.100	0.265	0.034	0.438	194.54	0.050
	5	0.600	0.500	0.500	0.169	-0.654	-0.117	82.03	0.102
	6	1.000	0.900	0.800	0.010	-2.078	-1.387	8.65	1.000
	7	0.100	0.000	0.000	0.717	-2.140	-1.265	7.57	0.803
	8	0.000	0.000	0.000	0.499	-2.094	-1.004	9.00	0.774
8	0	2.634	2.688	4196.9	0.002
	1	1.000	1.000	1.000	0.000	2.139	2.259	2362.0	0.004
	2	0.200	0.100	0.000	0.291	1.659	1.858	1332.55	0.007
	3	0.200	0.100	0.000	0.375	1.209	1.502	777.13	0.012
	4	0.300	0.200	0.200	0.357	0.783	1.188	480.87	0.022
	5	0.500	0.400	0.200	0.228	0.381	0.918	291.53	0.034
	6	0.300	0.200	0.200	0.309	-0.038	0.654	171.95	0.057
	7	0.500	0.300	0.000	0.181	-0.552	0.323	94.47	0.094
	8	0.900	0.900	0.900	0.016	-2.259	-1.168	9.00	1.000

p*: number of important variables. p: number of variables not including intercept

Table 3. Model selection methods when $\beta_j = 1/j$; $\sigma^2 = 0.1$, $n=10$, and $\rho=0.4$.

p*	p	powers			P-value	AIC	SIC	C _p	EFF
		$\alpha=0.10$	$\alpha=0.05$	$\alpha=0.01$					
0	0	-2.405	-2.349	6.90	1.000
	1	0.100	0.000	0.000	0.462	-2.682	-2.507	1.84	0.203
	2	0.000	0.000	0.000	0.982	-2.875	-2.489	0.61	0.137
	3	0.100	0.000	0.000	0.833	-3.076	-2.341	0.58	0.113
	4	0.000	0.000	0.000	0.811	-3.213	-1.910	1.69	0.103
	5	0.000	0.000	0.000	0.787	-3.226	-0.972	3.32	0.097
	6	0.000	0.000	0.000	0.833	-3.180	0.793	5.10	0.094
	7	0.000	0.000	0.000	0.907	-3.029	4.581	7.04	0.094
	8	0.000	0.000	0.000	0.734	-2.866	16.06	9.00	0.093
2	0	0.217	0.272	*****	0.056
	1	1.000	0.900	0.800	0.014	-1.532	-1.356	457244.75	0.352
	2	0.000	0.000	0.000	0.700	-2.321	-1.934	226922.00	0.945
	3	0.000	0.000	0.000	0.749	-2.623	-1.888	146607.95	0.703
	4	0.100	0.100	0.000	0.646	-2.941	-1.638	55040.66	0.570
	5	0.100	0.000	0.000	0.583	-3.211	-0.958	17355.46	0.529
	6	0.200	0.100	0.000	0.566	-3.813	0.159	3522.61	0.481
	7	0.100	0.100	0.000	0.628	-4.550	3.060	42.08	0.506
	8	0.000	0.000	0.000	0.673	-5.228	13.695	9.00	0.479
4	0	0.815	0.871	66417	0.046
	1	1.000	0.900	0.800	0.013	-0.746	-0.570	18101	0.240
	2	0.000	0.000	0.000	0.701	-1.589	-1.202	6144.3	0.517
	3	0.300	0.200	0.200	0.375	-2.409	-1.674	1690.3	0.844
	4	0.000	0.000	0.000	0.497	-2.928	-1.625	624.87	0.828
	5	0.100	0.100	0.000	0.716	-3.484	-1.230	119.67	0.784
	6	0.200	0.100	0.000	0.697	-3.856	0.117	22.92	0.778
	7	0.000	0.000	0.000	0.747	-4.083	3.527	11.78	0.750
	8	0.000	0.000	0.000	0.679	-4.618	14.31	9.00	0.767
6	0	0.919	0.975	4254.05	0.045
	1	1.000	0.800	0.500	0.030	-0.299	-0.123	1419.17	0.149
	2	0.100	0.100	0.100	0.490	-1.189	-0.802	827.43	0.402
	3	0.300	0.100	0.000	0.452	-1.919	-1.184	220.46	0.801
	4	0.000	0.000	0.000	0.672	-2.421	-1.119	54.98	0.791
	5	0.200	0.200	0.100	0.480	-3.343	-1.089	6.34	0.891
	6	0.000	0.000	0.000	0.766	-3.589	0.384	5.68	0.875
	7	0.000	0.000	0.000	0.808	-3.675	3.936	7.19	0.815
	8	0.000	0.000	0.000	0.754	-3.617	15.306	9.00	0.826
8	0	0.983	1.039	2085.6	0.042
	1	0.800	0.800	0.600	0.051	-0.191	-0.015	564.85	0.151
	2	0.100	0.100	0.000	0.450	-1.185	-0.798	126.73	0.364
	3	0.300	0.000	0.000	0.490	-1.638	-0.903	64.07	0.668
	4	0.000	0.000	0.000	0.568	-2.180	-0.877	26.70	0.822
	5	0.100	0.100	0.000	0.598	-2.765	-0.511	14.59	0.819
	6	0.100	0.100	0.000	0.578	-3.114	0.859	9.02	0.812
	7	0.000	0.000	0.000	0.848	-3.244	4.367	8.58	0.774
	8	0.000	0.000	0.000	0.659	-3.440	15.483	9.00	0.778

p*: number of important variables. p: number of variables not including intercept

Table 4. Model selection methods when $\beta_j = 1/j$, $\sigma^2 = 0.1$, $n=25$, and $\rho=0.4$.

p*	p	powers			P-value	AIC	SIC	Cp	EFF
		$\alpha=0.10$	$\alpha=0.05$	$\alpha=0.01$					
0	0	-2.530	-2.476	0.82	1.000
	1	0.000	0.000	0.000	0.495	-2.584	-2.464	-0.26	0.256
	2	0.000	0.000	0.000	1.000	-2.573	-2.375	0.29	0.215
	3	0.000	0.000	0.000	0.993	-2.555	-0.262	1.07	0.189
	4	0.000	0.000	0.000	0.896	-2.525	-2.120	2.18	0.176
	5	0.000	0.000	0.000	0.816	-2.482	-1.945	3.55	0.166
	6	0.000	0.000	0.000	0.771	-2.423	-1.731	5.20	0.161
	7	0.000	0.000	0.000	0.619	-2.354	-1.479	7.02	0.160
	8	0.000	0.000	0.000	0.652	-2.275	-1.184	9.00	0.160
2	0	0.563	0.617	404.77	0.006
	1	1.000	1.000	1.000	0.000	-1.135	-1.015	47.24	0.053
	2	0.100	0.100	0.100	0.723	-2.144	-1.945	3.08	1.000
	3	0.000	0.000	0.000	0.997	-2.205	-1.912	2.16	0.416
	4	0.000	0.000	0.000	0.877	-2.219	-1.814	2.23	0.310
	5	0.000	0.000	0.000	0.871	-2.180	-1.643	3.61	0.275
	6	0.000	0.000	0.000	0.778	-2.122	-1.430	5.24	0.257
	7	0.000	0.000	0.000	0.732	-2.052	-1.177	7.07	0.251
	8	0.000	0.000	0.000	0.573	-1.977	-0.886	9.00	0.248
4	0	0.806	0.860	619.95	0.017
	1	1.000	1.000	1.000	0.000	-0.297	-0.177	180.18	0.064
	2	0.400	0.300	0.100	0.363	-1.407	-1.208	35.26	0.226
	3	0.000	0.000	0.000	0.908	-1.856	-1.563	14.66	0.482
	4	0.300	0.300	0.200	0.512	-2.206	-1.801	4.17	0.920
	5	0.000	0.000	0.000	0.897	-2.214	-1.678	4.38	0.849
	6	0.000	0.000	0.000	0.837	-2.173	-1.481	5.68	0.794
	7	0.000	0.000	0.000	0.721	-2.120	-1.245	7.20	0.775
	8	0.000	0.000	0.000	0.623	-2.052	-0.962	9.00	0.782
6	0	0.937	0.992	628.26	0.013
	1	1.000	1.000	1.000	0.000	-0.230	-0.110	168.90	0.054
	2	0.300	0.200	0.100	0.617	-1.065	-0.866	57.30	0.149
	3	0.100	0.100	0.000	0.750	-1.625	-1.332	21.34	0.265
	4	0.100	0.000	0.000	0.628	-1.920	-1.515	11.05	0.403
	5	0.000	0.000	0.000	0.668	-2.116	-1.580	6.50	0.648
	6	0.100	0.000	0.000	0.670	-2.177	-1.485	5.86	0.988
	7	0.000	0.000	0.000	0.678	-2.132	-1.258	7.26	0.855
	8	0.000	0.000	0.000	0.560	-2.068	-0.978	9.00	0.902
8	0	0.905	0.959	449.59	0.014
	1	1.000	1.000	1.000	0.001	-0.070	0.050	144.38	0.041
	2	0.200	0.100	0.000	0.696	-0.732	-0.533	58.45	0.089
	3	0.100	0.000	0.000	0.632	-1.217	-0.924	27.28	0.167
	4	0.100	0.000	0.000	0.664	-1.532	-1.127	45.06	0.273
	5	0.000	0.000	0.000	0.589	-1.754	-1.218	9.10	0.442
	6	0.100	0.000	0.000	0.640	-1.858	-1.166	7.23	0.654
	7	0.000	0.000	0.000	0.681	-1.869	-0.994	7.57	0.841
	8	0.000	0.000	0.000	0.622	-1.823	-0.732	9.00	0.992

p*: number of important variables. p: number of variables not including intercept

Table 5. Model selection methods when $\beta_j=1$, $\sigma^2=0.1$, $n=10$, and $\rho=0.9$.

p*	p	powers			P-value	AIC	SIC	Cp	EFF
		$\alpha=0.10$	$\alpha=0.05$	$\alpha=0.01$					
0	0	-2.194	-2.139	1291.42	1.000
	1	0.200	0.200	0.100	0.373	-2.320	-2.144	1096.49	0.235
	2	0.000	0.000	0.000	0.667	-2.617	-2.231	874.41	0.111
	3	0.100	0.000	0.000	0.755	-2.746	-2.011	700.60	0.093
	4	0.000	0.000	0.000	0.690	-2.967	-1.665	469.69	0.082
	5	0.000	0.000	0.000	0.605	-3.230	0.976	215.64	0.077
	6	0.200	0.100	0.000	0.617	-3.714	0.259	47.04	0.074
	7	0.200	0.100	0.100	0.634	-4.117	3.494	23.60	0.072
	8	0.100	0.100	0.000	0.612	-4.473	14.450	9.00	0.071
2	0	1.420	1.476	82912.77	0.015
	1	1.000	1.000	1.000	0.000	-1.853	-1.677	1398.96	0.814
	2	0.100	0.000	0.000	0.742	-2.121	-1.734	1133.35	0.736
	3	0.200	0.000	0.000	0.631	-2.511	1.776	382.23	0.523
	4	0.100	0.100	0.100	0.609	-2.874	-1.572	298.99	0.543
	5	0.000	0.000	0.000	0.765	-3.048	-0.794	190.11	0.497
	6	0.100	0.100	0.000	0.448	-3.659	0.313	117.50	0.487
	7	0.100	0.000	0.000	0.704	-4.059	3.551	24.58	0.459
	8	0.100	0.000	0.000	0.639	-4.542	14.381	9.00	0.457
4	0	2.618	2.673	10688.25	0.005
	1	1.000	1.000	1.000	0.000	-1.263	1.088	336.19	0.284
	2	0.100	0.100	0.000	0.606	-2.199	-1.812	144.17	0.696
	3	0.000	0.000	0.000	0.890	-2.373	-1.638	105.04	0.847
	4	0.000	0.000	0.000	0.883	-2.451	1.149	80.54	0.795
	5	0.000	0.000	0.000	0.697	-2.725	-0.472	47.97	0.703
	6	0.100	0.000	0.000	0.780	-2.786	1.187	40.63	0.722
	7	0.000	0.000	0.000	0.814	-2.791	4.819	28.04	0.720
	8	0.100	0.100	0.000	0.715	-3.172	15.752	9.00	0.712
6	0	3.209	3.265	905789.38	0.004
	1	1.000	1.000	1.000	0.000	-0.601	0.462	10268.67	0.196
	2	0.200	0.100	0.100	0.353	-1.848	-1.462	1593.27	0.518
	3	0.100	0.100	0.100	0.598	-2.684	-1.949	534.05	0.749
	4	0.200	0.200	0.200	0.591	-3.397	-2.094	46.60	0.901
	5	0.000	0.000	0.000	0.736	-3.699	-1.445	29.99	0.908
	6	0.000	0.000	0.000	0.766	-3.947	0.026	12.92	0.919
	7	0.000	0.000	0.000	0.600	-4.299	3.311	9.90	0.877
	8	0.000	0.000	0.000	0.564	-4.867	14.056	9.00	0.862
8	0	3.770	3.825	49961.65	0.002
	1	1.000	1.000	1.000	0.000	-0.013	0.163	761.72	0.099
	2	0.300	0.300	0.100	0.399	-1.372	-0.985	241.32	0.337
	3	0.200	0.000	0.000	0.547	-2.262	-1.526	93.12	0.687
	4	0.000	0.000	0.000	0.673	-2.748	-1.445	33.04	0.858
	5	0.200	0.100	0.000	0.608	-3.309	-1.055	12.17	0.848
	6	0.000	0.000	0.000	0.742	-3.663	0.309	8.04	0.810
	7	0.000	0.000	0.000	0.754	-3.840	3.770	8.64	0.804
	8	0.100	0.100	0.000	0.666	-4.080	14.844	9.00	0.764

p*: number of important variables. p: number of variables not including intercept

Note that $k = \text{Rank}(X) = p + 1$.

Table 6. Model selection methods when $\beta_j = 1/j$, $\sigma^2 = 0.1$, $n=10$, and $\rho=0.9$.

p*	p	powers			P-value	AIC	SIC	Cp	EFF
		$\alpha=0.10$	$\alpha=0.05$	$\alpha=0.01$					
0	0	-2.167	-2.111	23.45	1.000
	1	0.100	0.100	0.100	0.351	-2.196	-2.020	19.32	0.325
	2	0.000	0.000	0.000	1.000	-2.386	-1.999	10.04	0.179
	3	0.000	0.000	0.000	0.794	-2.544	-1.809	8.26	0.146
	4	0.000	0.000	0.000	0.742	-2.733	-1.431	6.83	0.125
	5	0.000	0.000	0.000	0.734	-2.860	-0.606	6.11	0.115
	6	0.000	0.000	0.000	0.624	-3.242	0.731	5.23	0.105
	7	0.000	0.000	0.000	0.852	-3.143	4.467	7.10	0.104
	8	0.000	0.000	0.000	0.682	-3.029	15.894	9.00	0.103
2	0	0.708	0.764	10062.49	0.015
	1	1.000	1.000	1.000	0.000	-2.347	-2.171	562.83	0.810
	2	0.000	0.000	0.000	0.993	-2.539	-2.152	462.08	0.745
	3	0.100	0.000	0.000	0.773	-2.741	-2.005	216.93	0.592
	4	0.000	0.000	0.000	0.865	-2.806	-1.504	183.20	0.525
	5	0.000	0.000	0.000	0.731	-3.016	-0.463	154.10	0.470
	6	0.000	0.000	0.000	0.833	-2.997	0.975	139.73	0.442
	7	0.100	0.100	0.000	0.775	-3.311	4.299	10.22	0.420
	8	0.000	0.000	0.000	0.784	-3.510	15.413	9.00	0.414
4	0	1.183	1.239	1757.44	0.019
	1	1.000	1.000	1.000	0.000	-2.143	-1.968	43.19	0.623
	2	0.100	0.100	0.100	0.872	-2.412	-2.025	17.40	0.937
	3	0.300	0.100	0.000	0.551	-2.888	-2.153	11.04	0.481
	4	0.000	0.000	0.000	0.662	-3.049	-1.747	8.72	0.449
	5	0.100	0.100	0.100	0.811	-3.179	-0.925	6.69	0.438
	6	0.100	0.000	0.000	0.666	-3.433	0.548	6.11	0.433
	7	0.000	0.000	0.000	0.890	-3.386	4.224	7.54	0.430
	8	0.000	0.000	0.000	0.691	-3.499	15.424	9.00	0.424
6	0	1.513	1.569	7023.81	0.013
	1	1.000	1.000	1.000	0.000	-2.442	-2.266	66.95	0.650
	2	0.000	0.000	0.000	0.952	-2.727	-2.340	36.86	0.794
	3	0.000	0.000	0.000	0.692	-3.022	-2.287	22.34	0.745
	4	0.000	0.000	0.000	0.686	-3.287	-1.984	15.85	0.621
	5	0.100	0.000	0.000	0.578	-3.695	-1.441	11.22	0.576
	6	0.000	0.000	0.000	0.738	-3.818	0.155	10.32	0.562
	7	0.100	0.000	0.000	0.538	-4.365	3.246	7.92	0.537
	8	0.000	0.000	0.000	0.635	-4.649	14.274	9.00	0.528
8	0	1.686	1.742	2709.65	0.013
	1	1.000	1.000	1.000	0.000	-1.361	-1.456	103.82	0.670
	2	0.000	0.000	0.000	0.830	-2.287	-1.900	48.42	0.779
	3	0.000	0.000	0.000	0.925	-2.492	-1.757	36.80	0.693
	4	0.000	0.100	0.100	0.636	-2.963	-1.661	18.41	0.577
	5	0.100	0.100	0.000	0.562	-3.368	-1.114	11.50	0.533
	6	0.000	0.000	0.000	0.465	-3.494	0.479	10.36	0.509
	7	0.100	0.000	0.000	0.668	-3.678	3.932	8.20	0.501
	8	0.000	0.000	0.000	0.704	-3.830	15.093	9.00	0.499

p*: number of important variables. p: number of variables not including intercept

Table 7. Model selection methods when $\beta_j=1$, $\sigma^2=1$, $n=10$, and $\rho=0.9$.

p*	p	powers			P-value	AIC	SIC	Cp	EFF
		$\alpha=0.10$	$\alpha=0.05$	$\alpha=0.01$					
0	0	0.043	0.099	28560.45	1.000
	1	0.200	0.200	0.100	0.359	-0.111	0.064	18601.81	0.137
	2	0.000	0.000	0.000	0.584	-0.406	-0.020	8277.81	0.055
	3	0.000	0.000	0.000	0.823	-0.480	0.256	4072.16	0.046
	4	0.100	0.100	0.000	0.667	-0.684	0.619	888.92	0.041
	5	0.000	0.000	0.000	0.714	-0.791	1.463	327.00	0.038
	6	0.000	0.000	0.000	0.813	-0.805	3.168	253.22	0.035
	7	0.000	0.000	0.000	0.653	-0.978	6.633	114.94	0.035
	8	0.200	0.100	0.000	0.597	-1.758	17.165	9.00	0.034
2	0	1.657	1.713	134.45	0.083
	1	1.000	1.000	1.000	0.003	0.071	0.247	19.55	0.989
	2	0.000	0.000	0.000	0.768	-0.133	0.254	10.91	0.497
	3	0.100	0.100	0.100	0.554	-0.488	0.247	5.86	0.387
	4	0.100	0.000	0.000	0.648	-0.742	0.560	4.45	0.340
	5	0.000	0.000	0.000	0.807	-0.846	1.408	5.03	0.309
	6	0.000	0.000	0.000	0.702	-1.046	2.927	5.66	0.293
	7	0.000	0.000	0.000	0.737	-1.063	6.548	7.30	0.281
	8	0.000	0.000	0.000	0.738	-1.025	17.898	9.00	0.280
4	0	2.785	2.840	617.07	0.032
	1	1.000	1.000	1.000	0.000	0.015	0.191	25.20	0.983
	2	0.000	0.000	0.000	0.928	-0.056	0.331	19.24	0.744
	3	0.300	0.100	0.100	0.539	-0.588	0.147	7.27	0.556
	4	0.100	0.000	0.000	0.582	-0.854	0.449	5.00	0.507
	5	0.100	0.100	0.000	0.620	-1.103	1.151	5.17	0.468
	6	0.100	0.000	0.000	0.611	-1.370	2.603	5.98	0.430
	7	0.000	0.000	0.000	0.703	-1.581	6.030	7.12	0.412
	8	0.000	0.000	0.000	0.656	-1.492	17.431	9.00	0.409
6	0	3.199	3.255	802.90	0.029
	1	1.000	1.000	1.000	0.000	0.224	0.399	45.17	0.663
	2	0.000	0.000	0.000	0.933	-0.204	0.183	23.89	0.862
	3	0.000	0.000	0.000	0.811	-0.435	0.300	13.02	0.564
	4	0.000	0.000	0.000	0.826	-0.524	0.779	10.46	0.578
	5	0.000	0.000	0.000	0.753	-0.623	1.630	8.81	0.552
	6	0.000	0.000	0.000	0.709	-0.805	3.168	7.57	0.506
	7	0.000	0.000	0.000	0.664	-1.079	6.531	7.37	0.499
	8	0.000	0.000	0.000	0.749	-1.124	17.799	9.00	0.493
8	0	3.650	3.706	142877.50	0.018
	1	1.000	1.000	1.000	0.000	0.394	0.570	4459.69	0.451
	2	0.100	0.100	0.000	0.693	-0.313	0.074	1109.44	0.753
	3	0.000	0.000	0.000	0.890	-0.588	0.147	615.79	0.825
	4	0.000	0.000	0.000	0.752	-0.842	0.461	179.66	0.830
	5	0.100	0.100	0.000	0.820	-1.043	1.211	120.19	0.736
	6	0.000	0.000	0.000	0.842	-1.136	2.836	65.24	0.710
	7	0.000	0.000	0.000	0.659	-1.244	6.366	25.47	0.678
	8	0.100	0.100	0.000	0.647	-1.642	17.282	9.00	0.671

p*: number of important variables. p: number of variables not including intercept

비직교 회귀분석에서의 AIC , $SICc$, Cp 에 대한 연구

류 승 훈

응 용 수 학 과

한국해양대학교 대학원

요 약

본 논문에서는 독립변수가 비직교 일 때 최적합 회귀모형을 찾기 위해 AIC , $SICc$, Cp 의 검정력을 비교분석 하였다. 표본의 크기가 10과 25일 때 검정력과 p값을 찾고, 각 모형의 효율성을 모의 실험한다. 상관계수는 0.4와 0.9로 한다.

REFERENCES

- [1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in 2nd International symposium on Information Theory 267–281.(Eds) B.N. Petrov and F.Csaki, Akademia Kiado, Budapest.
- [2] Akaike, H.(1978) A bayesian analysis of the minimum AIC procedure. Annals of the Institute of Statistical Mathematics 30, Part A, 9–14.
- [3] Hurvich, C.M. and Tsai, C.L.(1989). Regression and time series model selection in small samples. *Bilmetrika* 76, 297–307.
- [4] Kullback, S. and Leibler, R.A.(1951) On information and sufficiency. *Annals of Mathematical Statistics* 22, 79–86.
- [5] Mallows, C.L.(1973). Some comments on Cp. *Thechnometrics*, 15, 661–675.
- [6] Mcquarrie, A.D.(1999). A small-sample correlation for the Schwarz SIC model selection criterion. *Statistics & Probability Letters* 44, 79–86.
- [7] Park, C.K.(2000) Restriction of Candidate Models in Orthogonal Regression, *The Journal of Korean Data Analysis Society*, Vol. 2, No.3, 377–388.
- [8] Park, C. K. and Park, C. I.(2001). Probabilities of Overfitting of Model Selection Criteria in Regression . *The Journal of Korean Data Analysis Society*, Vol.3, 245–253.
- [9] Park, C.K. and Yoda, T.(2000). Comparing of Model Selection Criteria in Regression. *OIKONOMIKA* Vol.36. No.3&4 93–106.

[10] Shibata, R.(1980). Asymptotically efficient selection of the order of the model for estimating parameters of linear process. *The Annals of Statistics* 8, 147-164.

[11] *STAT IMSL/LIBRARY* (1982). 2nd Version of Fortran Subroutines for Statistical Analysis.