Journal of the Research Institute of Industrial Technology. Vol. 23, 2006

Restriction of range of *SSE* to find the true model in regression

*Chan Keun Park**

**Division of Mathematical & Information Science and Semiconductor Physics, Korea Maritime University*

ABSTRACT : *In this paper, we introduce some model selection criteria and find their characters by reviewing the old papers. And we find the some model selection criteria of candidate models after we restrict the range of SSE in regression. We check the effect of extra variables using two models. In both models,* $n=25,000$, $Y=1+x+\epsilon$, *where* $x \sim N(0,1)$ and $\epsilon \sim N(0,1)$. *In model 1, there are two extra variables included and in model 2, there are five extra variables included in the study.*

KEYWORDS : Model Selection Criteria, *SSE*, F-distribution, $K-L$, L_2

1. Introduction

Selecting the best model in multiple regression has become a popular subject in recent years. Usually, we prefer to take simple model, but there are also important variables in the data. We should balance the simplicity and performance to find the best regression model. There are several steps to balance these two properties. First, check the relationship between variables and then select a good model from the set of candidate models. The model selection criterion is one of methods for selecting the best model. In this paper, we introduce some model selection criteria and find their characters by reviewing the old papers. And we find the some model selection criteria of candidate models after we restrict the range of regression models using *SSE*. *SSE* (Sum of Squared Error) is important thing to select the best model. But if we do not know the exact distribution of *SSE*, it is a little difficult to use *SSE* in model selection. We will control this difficulty at the further research. In this paper, we use the *SSE* when reduced model is nested and X is orthogonal matrix. In chapter 4, we calculate and compare the AIC, $AICc$, Cp and HQ after restriction of *SSE*. Also find the average *Kullback-Leibler* efficiency and the average L_2.

We first define the true regression model to be

$$Y = X_*\beta_* + \epsilon_* \qquad \epsilon_* \sim N(0, \sigma^2 I)$$

where $Y=(y_1, y_2, \ldots, y_n)'$ is an $n \times 1$ vector of responses, $X_*\beta_*$ is an $n \times 1$ vector of true unknown functions, and $\epsilon_* = (\epsilon_{*1}, \epsilon_{*2}, \ldots, \epsilon_{*n})'$. We assume that the errors ϵ_{*i} are independent and identically normally distributed, with constant variance σ_*^2 for $i = 1, 2, \ldots, n$.

We next define the general model to be

$$Y = X\beta + \epsilon \qquad \epsilon \sim N(0, \sigma^2 I)$$

where $X=(x_1, x_2, \ldots, x_n)'$ is a known $n \times k$ design matrix of rank k, x_i is a $k \times 1$ vector. β is a $k \times 1$ vector of unknown parameters, and $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$. We assume that the errors ϵ_i are independent and identically normally distributed with the constant variance σ^2 for $i = 1, 2, \ldots, n$. If the constant, or y-intercept, is included in the model, the first column of X will contain a column of 1's

* chapark@bada.hhu.ac.kr 051)410-4372

associated with constant.

Finally we will define the candidate model, with respect to the general model. In order to classify candidate model types we will partition X and β such that $X=(X_0,X_1,X_2)$ and $\beta=(\beta_0',\beta_1',\beta_2')'$, where X_0, X_1 and X_2 are $n\times k_0, n\times k_1$ and $n\times k_2$ matrices, and β_0, β_1 and β_2 are $k_0\times 1, k_1\times 1$ and $k_2\times 1$ vectors, respectively. If μ_* is a linear combination of unknown parameters such that $\mu_*=X_*\beta_*$, then underfitting will occur when $rank(X)<rank(X_*)$, and overfitting will occur when $rank(X_*)<rank(X)$. Thus we can rewrite the general model in the following form :

$$Y=X_0\beta_0+X_1\beta_1+X_2\beta_2+\epsilon$$
$$=X_*\beta_*+X_2\beta_2+\epsilon,$$

where $\beta_*=(\beta_0',\beta_1')'$, X_0 is the design matrix for an underfitted candidate model, $X_*=(X_0,X_1)$ is the design matrix for the true model and $X=(X_0,X_1,X_2)$ is the design matrix for an overfitted model. Thus an underfitted model is written as $Y=X_0\beta_0+\epsilon$ and an overfitted model is written as $Y=X_0\beta_0+X_1\beta_1+X_2\beta_2+\epsilon$.

There are two categories of model selection criteria. The goal is to select one model that best approximates the true model from a set of finite-dimensional candidate models. The candidate model that is closest to the true model is assumed to be the appropriate choice. Here, the term "closest" requires some well-defined distance or information measure in order to be evaluates. In large samples, model selection criteria that choose the model with minimum mean squared error distribution is said to be asymptotically efficient [11]. *AIC, AICc,* and *Cp* are asymptotically efficient criteria. We assume that the true model is of finite dimension, and that it is included in the set of candidate models. Under the this assumption the goal of model selection is to correctly choose the true model from the list of candidate models. A model selection criterion that identifies the correct model asymptotically with probability one is said to be consistent. *SIC* and *HQ* are consistent model selection criteria.

2. Historical Background of Literature

Much of past model selection research has been concerned with univariate or multiple regression models. Perhaps the first model selection criterion to be widely used is the adjusted R-squared, R^2_{adj}, which still appears in many regression texts today. It is known that R^2 always increases whenever a variable is added to the model, and therefore it will always recommend additional complexity without regard to relative contribution to model fit. R^2_{adj} attempts to correct for this always-increasing property. Other model selection work appeared in the late 60's and early 70's, most notably Akaike's FPE and Mallow's *Cp* [1][8]. The latter is currently one of the most commonly used model selection criteria for regression. Information theory approaches also appeared in th e1970's, which the landmark Akaike Information Criterion, based on the *Kullback-Leibler* discrepancy [2][3]. In the late 1970's there was an explosion of work in the information theory area, when the Bayesian Information Criterion (*BIC*), the Schwarz Information criterion (*SIC*), the Hannan and Quinn Criterion (*HQ*) [3][4][10]. Subsequently, in the late 1980's, Hurvich and Tsai adopted Sugiura's 1978 results to develope an improved small-sample unbiased estimator of the Kullback-Leibler discrepancy, *AICc* [8]. In 1980 the notion of asymptotic efficiency appeared in the literature as a paradigm for selecting the most appropriate model, and *SIC, HQ* became associated with the notion of consistency [9][11]. The model selection test, or say sequential F-test, is another model selection method. This method is applied by adding or deleting independent variables from the candidate model based on the F-test.

First of all, let's check the equations and characters of model selection criteria. Now, we review some common efficient criteria. Akaike showed that *AIC* is asymptotically unbiased for the *Kullback-Leibler* information up to a constant [1][7].

The distance measures L_2 and *Kullback-Leibler* discrepancy $(K-L)$ provide a way to evaluate how well the candidate model approximates the true model by estimating the difference between the expectation of

the vector Y under the true model and the candidate model. The L_2 distance between the estimated candidate model and the expectation of the true model can be defined as $L_2 = \frac{1}{n}\left\| X_*\beta_* - X\hat{\beta} \right\|^2$. Under the assumptions of normality, we define $(K-L)$ as $K-L = \frac{2}{n}E_*\left[\log\left(\frac{f_*}{f}\right)\right]$, where f_* and E_* denote the density and the expectation under the true model and f is the density function of the candidate mode [12]l.

Now we show that

$AIC = nlog(\hat{\sigma}_k^2) + 2(k+1) + nlog(2\pi) + n$ and the last two terms are not important for model selection, so we can ignore them. Simplifying and scaling 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 be overfitted in small samples [5]. Hurvich and Tsai attained the bias-corrected, in terms of selected order, version of *AIC*. Hurvich and Tsai modified *AIC* to provide an exactly unbiased estimator for the expected *K-L* information, assuming that the errors have a normal [5].

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

AIC and *AICc* are the methods related to *K-L* discrepancy [7]. The *K-L* discrepancy measures the distance of the density function of the true model and that of the candidate model. The minimum value of the *AIC* and *AICc* is said to be close to the true model. When the sample size is large, *AIC* and *AICc* behave the same. But, *AIC* has problems when the sample size is small so *AICc* is better than *AIC*. *AIC*, *AICc* and *Cp* are efficient model selection criteria [6].

SIC(BIC) can be overfitted in small samples due to the linear (in k) penalty function. The equation of *SIC* is

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

HQ is a strongly consistent estimation procedure based on the law of the iterated logarithm. The equation of *HQ* is

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

BIC and *HQ* are related to asymptotic performance properties. In small samples, say 25 observations or less, *BIC* tends to over fit the model. In large samples, *BIC* correctly identifies the true model. Such criteria are referred to us asymptotically consistent. If true model belongs to the set of candidate models and is of finite order, then a model selection that identifies the true model asymptotically with probability one is said to be consistent [11]. *HQ* is also a consistent criterion.

In Mallow's derivation, the estimate s_K^2 ($k=K$ in $s_k^2 = \frac{SSE_k}{n-k}$) from the largest candidate model was substituted as a potentially unbiased estimate of σ_*^2 to yield the well-known Mallows's *Cp* model selection criterion $Cp = \frac{SSE_k}{s_K^2} - n + 2k$.

The model selection test, or say sequential *F*-test, is another model selection method. It works well in practice but can't look at all possible subsets. So, it tends to have an underfitting problem. Terasvirta and Mellin compared the model selection and sequential *F*-test when all models are linear and one model is to be chosen from a set of nested alternative using Monte Carlo experiment [13]. They said that model selection tests seem to be well in simulation experiments.

3. Model Selection Criteria after Using *SSE*

Finding the "best" model with k variables is a little difficult, since there are 2^k possible models to check. We can compare *SSE* and number of parameters, k, of each subsets and then apply the model selection criteria to the list of k model. For the given parameter count k, the best model has the smallest *SSE*. In all subsets regression, the 2^k models reduce to a list of K models, the best for each parameter count, k=1,2,3,...,K. We need the distribution of *SSE* for comparing these models. To do that, develop the bootstrapping for all subsets regression, and then compare *SSE* for order k. Now, we check the one variable model, and find the best 1 variable model that

have the smallest *SSE* among the each 1 variable model. There should be k models to check the *SSE*. Find the best 2 variable model and there are ${}_kC_2$ models to check the *SSE*. We can keep going until the candidate model has k variables. Then, make the plot between *SSE* and best k models. Theoretically, this plot could be straight line, but in practice, the plot looks like curve. The *SSE* rapidly drops in some points and very slowly dropped in other points. Another possibility is to restrict range for existing of model selection criteria. This may prevent overfitting. We restrict the range of the subsets using plot. Bootstrap the distribution of the drop in *SSE* between SSE_k and SSE_{k+1}.

Now, let's talk about model selection using *SSE* of orthogonal case. Suppose the reduced model has m orthogonal remaining variables. And there are m nested full models each with 1 additional variable over the reduced model. The m full model could be $X_1, X_2, X_3, \ldots, X_m$. The SSE_r is the total SS and $SSE_r - SSE_{f1}$, $SSE_r - SSE_{f2}$, $\ldots$, $SSE_r - SSE_{fm}$ follow the $\sigma^2\chi^2$ distribution and they are independent each other. The proof of independence is below.

Theorem : $SSE_r - SSE_{f1}$ and $SSE_r - SSE_{f2}$ are independent when X is orthogonal and none of the new variables are shared.

proof) $SSE_R - SSE_{f1} = Y'[H_{f1} - H_r]Y$

$SSE_R - SSE_{f2} = Y'[H_{f2} - H_r]Y$

$(H_{f1} - H_r)(H_{f2} - H_r) = 0$

(By Fisher-Cochran Theorem)

So, we can say that all these chi-square distribution are independent since variables are orthogonal. The best full model is the model with largest chi-squared distribution with degree of freedom 1. Find the minimum of *SSE*. Since *SSE* of full model equals the *SSE* reduced model minus chi-square distribution, so finding the minimum of m independent chi-square distribution is the same as finding the minimum of *SSE*. We checked relationship between m and maximum of chi-square distribution when m is 50. We checked that the expected drop in *SSE* increases as m increases. When parameters increase over 26, the expected drop in *SSE* is increased very slowly. So, we can restrict the range of the subsets of parameters based on the drop. It means that we can reduce the selection of subsets of all regression models. And then we can apply the model selection criteria using these selected subsets.

We check the effect of extra variables using two models. In both models,
$n = 25{,}000$, $Y = 1 + X + \epsilon$, where $x \sim N(0,1)$, and $\epsilon \sim N(0,1)$. In model 1, there are two extra variables included in the study. In model 2, there are five extra variables included in the study. This example illustrates the impact of extra variables on selecting the true model. "$K-L$ *ave*" and "L_2 *ave*" denote the average *Kullback-Leibler* efficiency and the average L_2 efficiency, respectively, over 1,000 realizations. This chapter include theoretical properties of model selection criteria and the $K-L$ and L_2 distance. Here we use the expected values of L_2 and $K-L$ when discussing theoretical distance between the candidate model and true model. For L_2, we defined L_2 *expected efficiency* as

$$L_2 \ expected \ efficiency = \frac{E_{F_T}[L_2(M_c)]}{E_{F_T}[L_2(M_k)]}$$

where $E_{F_T}[L_2(M_c)]$ is the expected L_2 distance of the closest model and $E_{F_T}[L_2(M_k)]$ is the expected L_2 distance of the candidate model. Analogously, $K-L$ *expected efficiency* is defined as

$$K-L \ expected \ efficiency = \frac{E_{F_T}[K-L(M_c)]}{E_{F_T}[K-L(M_k)]}$$

where $E_{F_T}[K-L(M_c)]$ is the expected $K-L$ distance of the closest model and $E_{F_T}[K-L(M_k)]$ is the expected $K-L$ distance of the candidate model.

In the model 1, the model which has two variables is the best model and we added two more extra variables. Every model selection criteria chooses two variable model as the best model. The average *Kullback-Leibler* efficiency and average L_2 efficiency is close 1. It means that every criteria did well to find

the best model. In the model 2, also the two variable model is the best model and we added four extra variables. According to the above tables, the extra variables effect to choosing the more variables. That means it tends to overfitting if there are more extra variables in regression.

Model 1. Two Extra Variables Included in the True Model

k	*AIC*	*AICc*	*Cp*	*HQ*	*SIC*
1	0	0	0	0	0
2	691	691	691	924	994
3	268	268	268	74	6
4	41	41	41	2	0
true	692	691	691	924	994
K-L ave	0.815	0.815	0.815	0.947	0.995
L_2 *ave*	0.785	0.785	0.785	0.940	0.995

Model 2. : Five Extra Variables Included in the True Model

k	*AIC*	*AICc*	*Cp*	*HQ*	*SIC*
1	0	0	0	0	0
2	457	457	457	861	997
3	387	388	388	132	3
4	129	128	128	7	0
5	23	23	23	0	0
6	4	4	4	0	0
7	0	0	0	0	0
true	454	454	454	858	994
K-L ave	0.659	0.659	0.659	0.904	0.997
L_2 *ave*	0.611	0.611	0.611	0.893	0.997

4. Conclusions and Further Research

We reviewed some model selection criteria and compared these criteria when there are extra variables. If there are lots of variables in regression model, it's difficult to check the model selection criteria of every subset models. Before checking the all models, we had better to reduce the number of subset models using *SSE*. However, if m variables are orthogonal, *SSE* is followed the chi-square distribution. So, we can make compare *SSE* of every subsets using plot of chi-square distribution versus number of parameters.

However, if some new variables are shared in the full model, *SSE* is not independent. And if the m variables are not orthogonal, the distribution of *SSE* is so complicated we need to use bootstrapping. In the nested model, the distribution of *SSE* is known. In the non-nested model, distribution of *SSE* is unknown. In each case, need to develop the bootstrapping for all subsets regression.

In the next research, we will use the parametric bootstrapping method. First we assume a distribution $N(0, \hat{\sigma}^2)$ since we do not know the exact distribution of *SSE*. Then generate the residual and do the regression. In the basic F-test, $\frac{SSE_r - SSE_f}{SSE_f}$ is followed by F distribution and numerator and denominator are independent when these are followed the chi-squared distribution. In our research, we do not know th exact distribution of SSE and to develop the bootstrapping, we need to define the target function. Our proposed target function is :

$$F_* = \frac{1}{n-k_f} \times \frac{(SSE_r - SSE_f)}{SSE_f/(n-k_f)}$$

where k_f is the number of parameters in full model and n is the number of data. The target function is not exact F distribution since numerator and denominator are unknown distribution and nor independent each other. We need to show that this target function does not depend on σ^2 in nested and non-nested models.

References

[1] Akaike, H. 1969. Statistical predictor identification, *Annals of the Institute of Statistical Mathematics.* 22, 203-217.

[2] Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle, *In Second International Symposium on Information Theory*. Akademia Kiado, Budapest, 267-281.

[3] Akaike, H. 1978. A Bayesian analysis of the minimum *AIC* procedure. *Annals of the Institute of Statistical Mathematics.* 30, 9-13.

[4] Hannan, E.J. and B.G. Quinn 1979. The determination of the order of autoregression, *Journal of Royal Statistical Society B* 41, 190-265.

[5] Hurvich, C.M. and Tsai, C.L. 1989. Regression and time series model selection in small samples.

Biometrika 76, 297-307.

[6] Hurvich, C.M. and Tsai, C.L. 1990. Model selection for least absolute deviation regression in small samples, *Statistics &Probability letters* 9, 259-265.

[7] Kullback, S. and Leibler, R.A. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22, 79-86.

[8] Mallows, C.L. 1973. Some comments on *Cp*, *Technometrics* 12, 591-612.

[9] Nishii, R. 1984. Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics* 12, 758-765.

[10] Schwartz, G. 1978. Estimating the dimension of a model, *Annals of Statistics* 6, 461-464.

[11] Shibata, R. 1980. Asymptotic efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics* 8, 147-164.

[12] Sugiura, N. 1992. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics* 7, 275-277.

[13] Terasvira, T. and Mellin, I. 1986. Model selection criteria and model selection tests in regression models, *Scandinavian Journal of Statistics* 13, 159-171.

Received : 24 November 2005
Accepted : 10 January 2006