The Korean Communications
in Statistics Vol. 4, No. 1, 1997
pp. 11-19

# Comparisons Between Model Selection Criteria

**Choongrak Kim[1], Hyoungsoon Kim[2], Meeseon Jeong[3]**

## Abstract

One of the most important issues in regression is variable selection problem. Recently several methods have been proposed to overcome the overparameterization property of Mallow's $C_p$. In this paper we compare these model selection criteria in view of the performance of selecting true model by simulation study.

## 1. Introduction

One of the most important fields in regression analysis is subset selection. Subset selection aims at two targets, variance reduction and model simplicity. There are various methods for selecting good subsets of variables. Usually, a regression equation based on a few variables will be more accurate and certainly simpler than the equation based on all candidate variables. See Breiman (1995) for detailed discussion.

One of the most popular criterion is $C_p$ by Mallows (1973). However, it is well-known that $C_p$ tends to select an unnecessarily large model (Miller 1984, 1990; Breiman 1992; Shao 1993). Other methods such as the Akaike information criterion (Akaike 1974), the jackknife, the cross-validation, and the bootstrap (Efron 1983, 1986) are asymptotically equivalent to $C_p$ (Stone 1977; Efron 1983). Recently several methods overcoming the problem of overparameterization are proposed; multifold cross-validation (MCV) by Burman (1989), Zhang (1993), and Shao (1993), little bootstrap (LB) by Breiman (1992), non-negative garrote (NG) by Breiman (1995), and $K_p$ by Kim (1996).

In this paper we compare the performance of selecting true model in $C_p$, MCV, LB, NG, and $K_p$. Summary on these criteria is given in Section 2. In Section 3 comparison of computation time is done and we investigate performance of selecting true model via Monte Carlo studies. Concluding remarks are given in Section 4.

1) Associate Professor, Department of Statistics, Pusan National University, Pusan, 609-735, Korea
Member of Research Institute for Computer, Information and Communication, Pusan National University
2) Department of Statistics, Pusan National University, Pusan, 609-735, Korea

## 2. Model Selection Criteria

Consider a linear model

$$y = X\beta + \varepsilon \tag{2.1}$$

where $y$ is an $n$-vector of response, $X$ is an $n \times k$ design matrix with 1's in the first column, $\beta$ is a $k$-vector of unknown coefficients, and $\varepsilon$ is error terms with $E(\varepsilon) = 0$ and $Cov(\varepsilon) = \sigma^2 I$. We assume that the design matrix $X$ contains all candidate predictors, so (2.1) is often called full model. By parsimony, it is required to select small numbers of predictors based on a given criterion. That is, given the data form $\{(y_i, x_{1i}, x_{2i}, \cdots, x_{ki}),\ i = 1, \cdots, n\}$, some of the predictor variables $\mathbf{x}_1, \cdots, \mathbf{x}_k$ are eliminated and the prediction equation for $y$ is based on the remaining set of variables.

$C_p$ is aimed to minimize the mean squared error, and if the current model consists of $p-1\ (<k)$ predictors $C_p$ is given by

$$C_p = \frac{\sum e_i^2}{s^2} - (n - 2p) \tag{2.2}$$

where $e_i$'s are residuals from the current model and $s^2$ is unbiased estimator of $\sigma^2$ from the full model (2.1). Therefore, it is required to evaluate $\sum_{i=1}^{k-1} \binom{k-1}{i}$ times of $C_p$ to select a subset minimizing $C_p$.

MCV has been suggested by Burman (1989), Zhang (1993), and Shao (1993). While the original CV is leave-one-out cross-validation, MCV is leave-$m$-out ($m>1$) cross-validation. In this paper, we present a version of MCV by Shao (1993). Suppose that $n$ data points are available for selecting a model from a class of models. The data set is split into two parts. The first part contained $n-m$ data points used for model construction, whereas the second part contains $m$ data points reserved for model validation. The size of $m$ should satisfy $m/n \to 1$. To reduce the computational amount, several techniques were suggested; balanced incomplete block design, Monte Carlo, and analytic approximation. Shao (1993) concluded that Monte Carlo technique is most practical and efficient. To be more specific, split the data set into two parts: $\{(y_i, \mathbf{x}_i),\ i \in Q\}$ and $\{(y_i, \mathbf{x}_i),\ i \in Q^c\}$, where $Q$ is a subset of $\{1, \cdots, n\}$ containing $m$ integers and $Q^c$ is its complement containing $n-m$ integers. Then randomly draw a collection $R$ of $b$ subsets of size $m$ and select a model by minimizing

$$MCV = \frac{1}{mb} \sum_{Q \in R} \| y_Q - \hat{y}_{(Q)} \|^2 \tag{2.3}$$

where $y_Q$ is the $m$-vector containing the components of $y$ indexed by $i \in Q$ and $\hat{y}_{(Q)}$ is the prediction of $y_Q$ using $\mathbf{x}_i$'s with $i \notin Q$. Therefore, MCV in (2.3) is a general version of CV, i.e., MCV reduces to CV if $m = 1$. Shao (1993) insisted that $b \geq n$ is enough, and used $b = 2n$ in his simulation study. Note that MCV requires $mb$ times more than CV or $C_p$.

Breiman (1992) suggested a variable selection technique called the little bootstrap. This approach is through the criterion that a good selection procedure selects dimensionality so as to give low prediction error which is defined as follows :
Suppose that the true model is

$$y_i = \mu(\mathbf{x}_i) + \varepsilon_i$$

and consider generating new data of the form

$$y_i^{new} = \mu(\mathbf{x}_i) + \varepsilon_i^{new}$$

with the $\{\varepsilon_i^{new}\}$ independent of the $\{\varepsilon_i\}$ but having the same distribution. Then, the prediction error (PE) is defined as

$$PE = E \sum_i (y_i^{new} - \mathbf{x}_i \beta)^2$$

where the expectation is over $\{y_i^{new}\}$. LB estimate of PE of the current model is obtained by the following steps;

1. Generate $\{\varepsilon_{1i}\}$, $i = 1, \cdots, n$ as i.i.d. $N(0, t^2 s^2)$ for some $t$ and form the new $y$ data

$$\tilde{y} = y + \varepsilon_1 .$$

2. Calculate

$$\frac{1}{t^2} \sum_{i=1}^{n} \varepsilon_{1i} (\hat{y}_i - \hat{\tilde{y}}_i)$$

where $\hat{y}_i$ is a fitted value of $y_i$ from the full model and $\hat{\tilde{y}}_i$ is a fitted value of $\tilde{y}_i$ from the current model.

3. Repeat step 1 and 2 a number of times and denote this average $B_t$.

4. The LB estimate is

$$\sum e_i^2 - ns^2 - 2B_t .$$

Throughout simulation experiments, Breiman (1992) proposed that the best range for $t$ in step 1 and 2 is $[0.6, 0.8]$ and that averaging over 25 repetitions to form $B_t$ in step 3 is usually sufficient. By this procedure a model is selected with the minimum LB estimate.

As a new model selection criterion Breiman (1995) propose non-negative garrote. Let $\{\hat{\beta}_j\}$ be the original OLS estimates based on the full model. Take $\{c_j\}$ to minimize

$$\sum_i \left( y_i - \sum_j c_j \hat{\beta}_j x_{ji} \right)^2 \tag{2.4}$$

under the constraints

$$c_j \geq 0, \quad \sum_j c_j \leq g, \quad j = 1, \cdots, k \tag{2.5}$$

where $g$ is called the garrote parameter.

Then, Breiman (1995) suggested $\tilde{\beta}_j = c_j \hat{\beta}_j$ as the new coefficient estimate. The procedure in (2.4) and (2.5) is in fact the constrained least squares minimization problem. A Fortran subroutine that outputs the values of the $\{c_j\}$ for any value of $g$, $0 < g < k$, is available by ftp to stat-ftp.berkeley.edu in the directory /pub/user/breiman.

Kim (1996) suggested a model selection criterion $K_p$ by the replacement measure. The idea is based on the robust estimation in regression. The robust regression estimation is done by pulling observations towards their fitted values, and refitting iteratively until convergence is obtained, i.e., if a current model is appropriate then estimates based on the replaced response are very close to estimates based on the original data. In fact, $K_p$ is based on the replacement measure $R_i$ which is sum of these differences and has interesting connection with influence measures like the influence curve and the local influence by Cook (1986). If the model is robust, then replacing $y_i$ by $\hat{y}_i$ for each $i$ does not alter $\hat{\beta}$ seriously, i.e., $\hat{\beta} - \hat{\beta}_{r(i)}$, where $\hat{\beta} = (X'X)^{-1}X'y$ and $\hat{\beta}_{r(i)} = (X'X)^{-1}X'y_{r(i)}$, $y_{r(i)} = (y_1, \cdots, y_{i-1}, \hat{y}_i, y_{i+1}, \cdots, y_n)$, would be small. Define the effect of replacement by a scalar version of $\hat{\beta} - \hat{\beta}_{r(i)}$ as

$$R_i = (\hat{\beta} - \hat{\beta}_{r(i)})' X'X (\hat{\beta} - \hat{\beta}_{r(i)}) / ps^2$$

by mimicking the Cook's distance (Cook (1977)). It can be shown that

$$R_i = e_i^2 h_{ii} / ps^2$$

and $K_p$ is given by

$$K_p = n \sum R_i - (n - 2p). \tag{2.6}$$

Therefore, $K_p$ is sum of the effect of replacing $y_i$ by $\hat{y}_i$. If the current model is appropriate $\sum e_i^2 h_{ii}$ has expected value $p(n-p)s^2/n$ approximately since $h_{ii}$ is near $p/n$. Therefore, approximately, $K_p$ has expected value $p$ if the current model is appropriate.

# 3. Simulation Study

## 3.1 Computation Time

To compare computation times of $C_p$, MCV, LB, NG, and $K_p$, consider the same model as Shao(1993):

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i$$

where $i = 1, \cdots, n$. Each $x_{ki}$, $k = 1,2,3,4$, is generated from $U(0,1)$, and $\varepsilon_i$ are i.i.d. from $N(0,1)$. Also, $\beta = (2,6,4,0,0)$ is specified, and the simulation is done for $n = 20,40,60,80,100$.

Computation times of $C_p$, MCV, LB, NG, and $K_p$ are plotted in Figure 1. Random number generations are done by IMSL, and the computation times are obtained by "TIMDY" command in Fortran. The results are summarized as follows. $K_p$ and $C_p$ take almost the same computation times. And LB and MCV require much more computation time than $C_p$ or $K_p$, NG requires several times compared other criteria, and increases in an exponential rate as $n$ increases. This gap will be wider if $n$ become large.

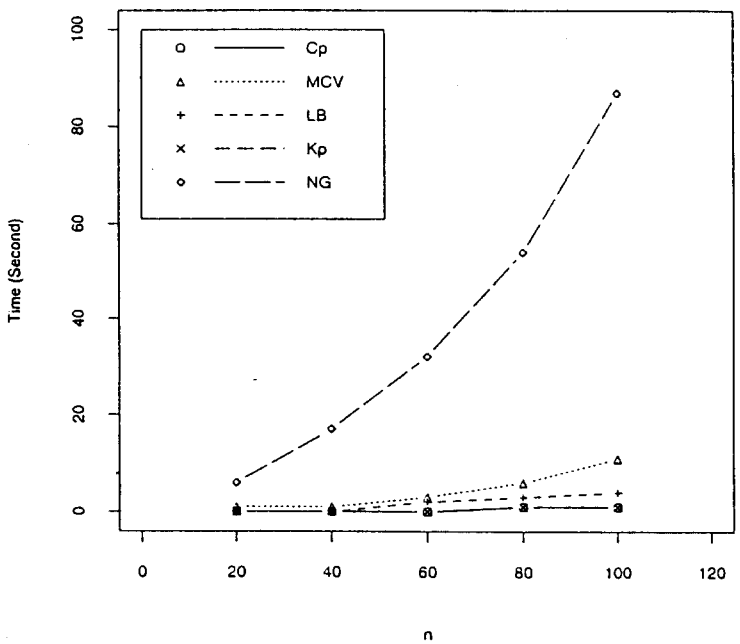


**Figure 1**: Plot of computation times of $C_p$, MCV, LB, NG, and $K_p$ for $n = 20,40,60,80,100$ (unit of time is second)

### 3.2 Performance of Selecting a True Model

To see the finite sample performance of selecting a true model of $C_p$, MCV, LB, NG, and $K_p$, a simulation study is done. Consider the same model as Shao (1993) used:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

where $i = 1, \cdots, 40$, $\varepsilon_i$ are i.i.d. from $N(0,1)$, $x_{ki}$ is the $i$th value of the $k$th predictor, and the values of $x_{ki}$ are taken from Shao (1993), which is listed in Table 1. Let $\beta_1 = 2$ and $x_{1i} \equiv 1$. Some of the $\beta_k$ may be 0. So, some predictors are selected from five possible variables $\{ x_1, \cdots, x_5 \}$, the model with the best predictive ability is chosen. To see the performance of selecting true model for each criterion, we compute the number of selecting true model out of 1,000 replications.

**Table 1**: Data set in Shao (1993)

| $i$ | $x_{2i}$ | $x_{3i}$ | $x_{4i}$ | $x_{5i}$ | $i$ | $x_{2i}$ | $x_{3i}$ | $x_{4i}$ | $x_{5i}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | .36 | .53 | 1.06 | .5326 | 21 | .09 | .18 | .59 | .1855 |
| 2 | 1.32 | 2.52 | 5.74 | 3.6183 | 22 | .02 | .16 | .24 | .1572 |
| 3 | .06 | .09 | .27 | .2594 | 23 | .02 | .11 | .21 | .0998 |
| 4 | .16 | .41 | .83 | 1.0346 | 24 | .05 | .24 | .43 | .2804 |
| 5 | .01 | .02 | .07 | .0381 | 25 | .11 | .39 | .29 | .2879 |
| 6 | .02 | .07 | .07 | .3440 | 26 | .18 | .11 | .43 | .6810 |
| 7 | .56 | .62 | 2.12 | 1.4559 | 27 | .04 | .09 | .23 | .3242 |
| 8 | .98 | 1.06 | 2.89 | 4.0182 | 28 | .85 | 1.33 | 2.70 | 2.6013 |
| 9 | .32 | .20 | .76 | .4600 | 29 | .17 | .32 | .66 | .4469 |
| 10 | .01 | .00 | .07 | .1540 | 30 | .08 | .12 | .49 | .2436 |
| 11 | .15 | .25 | .50 | .6516 | 31 | .38 | .18 | .49 | .4400 |
| 12 | .24 | .28 | .59 | .0611 | 32 | .11 | .13 | .18 | .3351 |
| 13 | .11 | .35 | .40 | .1922 | 33 | .39 | .38 | .99 | 1.3979 |
| 14 | .08 | .13 | .28 | .0931 | 34 | .43 | .46 | 1.47 | 2.0138 |
| 15 | .61 | .85 | .49 | .0538 | 35 | .57 | 1.16 | 1.82 | 1.9356 |
| 16 | .03 | .03 | .23 | .0199 | 36 | .13 | .03 | .08 | .1050 |
| 17 | .06 | .11 | .50 | .0419 | 37 | .04 | .05 | .14 | .2207 |
| 18 | .02 | .08 | .25 | .1093 | 38 | .13 | .18 | .28 | .0180 |
| 19 | .04 | .24 | .08 | .0328 | 39 | .20 | .95 | .41 | .1017 |
| 20 | .00 | .02 | .04 | .0797 | 40 | .07 | .06 | .18 | .0962 |

We used $t = 0.6$ in step 2 of LB procedure. In step 3 of LB procedure Breiman (1992) used 40 repetitions, but we repeat 250 repetitions for more accurate result. Table 2 gives the empirical results based on 1,000 replications of selecting each model in several different cases. The simulation results in table 2 can be summarized as follows:

1. Overparameterization of $C_p$ is clear, and LB shows similar trend although it is more severe than $C_p$.
2. In most cases, MCV and $K_p$ outperform $C_p$ or LB. Also, $K_p$ is slightly better than MCV except the largest model case.
3. NG outperforms $C_p$, LB, or MCV, and it is comparable to $K_p$.

**Table 2**: Numbers of selecting a true model based on 1,000 replications

| $\beta'$ | Model | $C_p$ | LB | MCV | $K_p$ | NG |
|---|---|---|---|---|---|---|
| (2,0,0,4,0) | 1,4 (true) | 593 | 499 | 816 | 834 | 842 |
| | 1,2,4 | 97 | 127 | 69 | 38 | 0 |
| | 1,3,4 | 112 | 126 | 86 | 18 | 0 |
| | 1,4,5 | 116 | 125 | 22 | 97 | 0 |
| | 1,2,3,4 | 35 | 42 | 6 | 7 | 0 |
| | 1,2,4,5 | 20 | 28 | 0 | 0 | 0 |
| | 1,3,4,5 | 20 | 36 | 1 | 4 | 0 |
| | 1,2,3,4,5 | 7 | 17 | 0 | 2 | 158 |
| (2,0,0,4,8) | 1,4,5 (true) | 723 | 575 | 785 | 931 | 868 |
| | 1,2,4,5 | 114 | 154 | 102 | 39 | 0 |
| | 1,3,4,5 | 127 | 201 | 108 | 16 | 0 |
| | 1,2,3,4,5 | 36 | 70 | 5 | 14 | 132 |
| (2,9,0,4,8) | 1,4,5 | 0 | 0 | 23 | 90 | 0 |
| | 1,2,4,5 (true) | 830 | 759 | 871 | 886 | 851 |
| | 1,3,4,5 | 0 | 0 | 7 | 2 | 0 |
| | 1,2,3,4,5 | 170 | 241 | 99 | 22 | 149 |
| (2,9,6,4,8) | 1,2,3,5 | 0 | 0 | 6 | 0 | 1 |
| | 1,2,4,5 | 0 | 0 | 12 | 55 | 0 |
| | 1,3,4,5 | 1 | 0 | 55 | 30 | 6 |
| | 1,2,3,4,5 (true) | 999 | 1000 | 927 | 915 | 993 |

## 4. Concluding Remarks

The issue of subset selection is very important in statistical regression area. One of the most popular criterion in this problem is $C_p$. However, many authors argued that the $C_p$ tends to select an unnecessarily large model. In this paper, we compare the performance of selecting true model in $C_p$, MCV, LB, NG, and $K_p$. Throughout a simulation study, MCV and $K_p$ outperform $C_p$ or LB, and the performance of NG is as good as $K_p$. We have given evidence that the NG is a worthy competitor to subset selection methods. It provides simple regression equation with better predictive accuracy. However, NG requires a lot of computation times.

## References

[1] Akaike, M. (1974). A New Look at Statistical Model Identification, *IEEE Transactions on Automatic Control*, Vol. 19, 716-723.

[2] Breiman, L. (1992). The Little Bootstrap and Other Methods for Dimensionality Selection in Regression : X-Fixed Prediction Error, *Journal of the American Statistical Association*, Vol. 87, 738-754.

[3] Breiman, L. (1995). Better Subset Regression Using the Nonnegative Garrote, *Technometrics*, Vol. 37, 373-384.

[4] Burman, P. (1989). A Comparative Study of Ordinary Cross-Validation, $v$-Fold Cross-Validation, and the Repeated Learning-Testing Methods, *Biometrika*, Vol. 76, 503-514.

[5] Cook, R. D. (1977). Detection of Influential Observations in Linear Regression, *Technometrics*, Vol. 19, 15-18.

[6] Cook, R. D. (1986). Assessment of Local Influence(with discussion), *Journal of the Royal Statistical Society, Ser. B.* Vol. 48, 133-169.

[7] Efron, B. (1983). Estimating the Error Rate of a Prediction Rule : Improvement on Cross-Validation, *Journal of the American Statistical Association*, Vol. 78, 316-331.

[8] Efron, B. (1986). How Biased is the Apparent Error Rate of a Prediction Rule? *Journal of the American Statistical Association*, Vol. 81, 461-470.

[9] Kim, C. (1996). Local Influence and Replacement Measure, *Communications in Statistics-Theory and Methods*, Vol. 25, 49-61.

[10] Mallows, C. L. (1973). Some Comments on $C_p$, *Technometrics*, Vol. 15, 661-675.

[11] Miller, A. J. (1984). Selection of Subsets of Regression Variables(with discussion), *Journal of the Royal Statistical Society, Ser. A.* Vol. 147, 389-425.

[12] Miller, A. J. (1990). *Subset Selection in Regression,* London: Chapman and Hall.

[13] Shao, M. (1993). Linear Model Selection by Cross-Validation, *Journal of the American Statistical Association*, Vol. 88, 486-494.

[14] Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion, *Journal of the Royal Statistical Society, Ser. B.* Vol. 39, 44-47.

[15] Zhang, P. (1993). Model Selection via Multifold Cross Validation, *The Annals of Statistics*, Vol. 21, 299-313.