An Unbiased C_p Criterion for Multivariate Ridge Regression

(Last Modified: March 7, 2008)

Hirokazu YANAGIHARA[1] AND Kenichi SATOH[2]

[1]*Department of Mathematics, Graduate School of Science, Hiroshima University*
1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8626, Japan

[2]*Department of Environmetrics and Biometrics*
Research Institute for Radiation Biology and Medicine, Hiroshima University
1-2-3 Kasumi, Minami-ku, Hiroshima, Hiroshima 734-8553, Japan

Abstract

Mallows' C_p statistic is widely used for selecting multivariate linear regression models. It can be considered to be an estimator of a risk function based on an expected standardized mean square error of prediction. Fujikoshi and Satoh (1997) have proposed an unbiased C_p criterion (called modified C_p; MC_p) for selecting multivariate linear regression models. In this paper, the unbiased C_p criterion is extended to the case of a multivariate ridge regression model. It is analytically proved that the proposed criterion has not only smaller bias but also smaller variance than an existing C_p criterion, and we show that our criterion has useful properties by means of numerical experiments.

AMS 2000 *subject classifications*: Primary 62J07; Secondary 62F07.
Key words: Bias correction; Mallows' C_p statistic; Model selection; Multivariate linear regression model; Ridge regression.

1. Introduction

Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$ be an $n \times p$ observation matrix and $\boldsymbol{X}$ be an $n \times k$ matrix of explanatory variables of full rank k. Suppose that j denotes a subset of $\omega = \{1, \ldots, k\}$ containing k_j elements, and let $\boldsymbol{X}_j$ denote the $n \times k_j$ matrix consisting of the columns of $\boldsymbol{X}$ indexed by the elements of j. Then we consider the following candidate model with k_j explanatory variables:

$$\boldsymbol{Y} \sim N_{n \times p}(\boldsymbol{X}_j \boldsymbol{\Xi}_j, \boldsymbol{\Sigma}_j \otimes \boldsymbol{I}_n). \tag{1.1}$$

[1]Corresponding author, E-mail: *yanagi@math.sci.hiroshima-u.ac.jp*

We call the model with $\boldsymbol{X}_\omega = \boldsymbol{X}$ the full model. Then we estimate $\boldsymbol{\Xi}_j$ by ridge-regression, i.e.,

$$\hat{\boldsymbol{\Xi}}_{j,\theta} = \boldsymbol{M}_{j,\theta}^{-1}\boldsymbol{X}_j'\boldsymbol{Y}, \tag{1.2}$$

where $\boldsymbol{M}_{j,\theta} = \boldsymbol{X}_j'\boldsymbol{X}_j + \theta\boldsymbol{I}_{k_j}$ $(\theta \geq 0)$. Notice that $\hat{\boldsymbol{\Xi}}_{j,0}$ is the ordinary maximum likelihood estimator of $\boldsymbol{\Xi}_j$ (or the ordinary least square estimator of $\boldsymbol{\Xi}_j$). In the above situation, optimization of the subset j and the ridge parameter θ is an important problem.

Choosing optimal j and θ so as to minimize a risk function is very well known method for model selection. In this paper, we consider the expected mean square error (MSE) of prediction as a risk function. It measures the discrepancy between a predictor of $\boldsymbol{Y}$ and a future observation, or imaginary new observation. However, we cannot directly use such a risk function in a real situation, because it includes unknown parameters. In practice, we use an estimator that is an information criterion, instead of the risk function. Obtaining an unbiased estimator of the risk function will allow us to correctly evaluate the discrepancy between the predictor of $\boldsymbol{Y}$ and a future observation, which will further facilitate the selection of optimal j and θ.

In this paper, we call an estimator of the risk function, based on the expected MSE of prediction, a C_p criterion, because Mallows' C_p statistic (Mallows, 1973; 1995) can be considered to be an estimator of such a risk when the candidate models are univariate linear regression models. When an observation is univariate, the discrepancy used consists of the Euclidean distance between the predictor and the future observation. However, when observation is multivariate, we need to take into account the correlation between response variables. Hence we have to use the discrepancy based on the Mahalanobis distance between them, i.e., the expected MSE standardized by the true variance-covariance matrix of observation. Such a risk function was proposed by Fujikoshi and Satoh (1997). Since the true variance-covariance matrix is unknown, we must replace it by its estimator. This replacement makes it hard to obtain an unbiased C_p criterion, because the estimated regression coefficient matrix and the estimated variance-covariance matrix are not independent, making this case more difficult to handle than the case of a multivariate linear regression model. Nevertheless, we can develop an unbiased C_p criterion even for the multivariate ridge-regression model by decomposing the residuals sum of squares into two parts, where the first part depends on the estimated variance-covariance matrix and the other part is independent of the estimated variance-covariance matrix. Such a

decomposition can be derived from the formula in Draper and Herzberg (1987). The definition of our unbiased C_p criterion is very simple, and it is not necessary to carry out complicated calculus to obtain an unbiased criterion, such as in Hurvich, Simonoff and Tsai (1998). Moreover, we are able to prove analytically that the proposed criterion has not only smaller bias but also smaller variance than the existing C_p criterion. We call it the modified C_p (MC_p) criterion, because our unbiased C_p coincides with the criterion in Fujikoshi and Satoh (1997) when the ridge parameter is 0.

This paper is organized in the following way: In Section 2, we propose the MC_p criterion for the multivariate ridge regression model by using the formula in Draper and Herzberg (1987). Several mathematical properties of our criterion are shown in Section 3. In Section 4, we examine the performance of the proposed criterion by conducting numerical simulations. Section 5 contains a discussion and our conclusions. Technical details are provided in the Appendix.

2. Unbiased C_p Criterion

Suppose that the true model of $\boldsymbol{Y}$ is expressed as

$$\boldsymbol{Y} \sim N_{n\times p}(\boldsymbol{\Gamma}_*, \boldsymbol{\Sigma}_* \otimes \boldsymbol{I}_n). \tag{2.1}$$

Let $\boldsymbol{P}_{\boldsymbol{A}}$ be the projection matrix to the subspace spanned by the columns of $\boldsymbol{A}$, i.e., $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$. Then, we suppose that the following assumption is satisfied.

- Assumption: at least the full model includes the true model, i.e., $\boldsymbol{P}_{\boldsymbol{X}_\omega}\boldsymbol{\Gamma}_* = \boldsymbol{\Gamma}_*$.

Let $\hat{\boldsymbol{Y}}_{j,\theta}$ be the predictor of $\boldsymbol{Y}$ given by $\hat{\boldsymbol{Y}}_{j,\theta} = \boldsymbol{X}_j\hat{\boldsymbol{\Xi}}_{j,\theta}$ and $\boldsymbol{U}$ be an $n\times p$ random variable matrix which is independent of $\boldsymbol{Y}$ and has the same distribution as $\boldsymbol{Y}$. $\boldsymbol{U}$ is regarded as a future observation or imaginary new observation. As a criterion for the goodness of fit of the candidate model, we consider the underlying risk function based on the MSE of prediction, which is proposed by Fujikoshi & Satoh (1997).

$$R(j,\theta) = E^*_{\boldsymbol{Y}}E^*_{\boldsymbol{U}}\left[\mathrm{tr}\left\{(\boldsymbol{U}-\hat{\boldsymbol{Y}}_{j,\theta})\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{U}-\hat{\boldsymbol{Y}}_{j,\theta})'\right\}\right],$$

where E^* denotes the expectation under the true model in (2.1). We regard the model with $j^{(r)}$ and $\theta^{(r)}$ which minimizes $R(j,\theta)$ as the principal best model. Let $\boldsymbol{W}_{j,\theta}$ be the residual matrix for the ridge regression, i.e.,

$$\boldsymbol{W}_{j,\theta} = (\boldsymbol{Y}-\hat{\boldsymbol{Y}}_{j,\theta})'(\boldsymbol{Y}-\hat{\boldsymbol{Y}}_{j,\theta}) = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{X}_j\boldsymbol{M}_{j,\theta}^{-1}\boldsymbol{X}_j')^2\boldsymbol{Y}. \tag{2.2}$$

By simple calculation, $R(j,\theta)$ can be rewritten as

$$R(j,\theta) = E^*_{\boldsymbol{Y}}\left[\mathrm{tr}(\boldsymbol{W}_{j,\theta}\boldsymbol{\Sigma}_*^{-1})\right] + 2p\mathrm{tr}(\boldsymbol{M}_{j,\theta}^{-1}\boldsymbol{M}_{j,0}). \tag{2.3}$$

Therefore we can propose a rough estimator for the risk function by using an estimator for $E^*_{\boldsymbol{Y}}[\mathrm{tr}(\boldsymbol{W}_{j,\theta}\boldsymbol{\Sigma}_*^{-1})]$.

Let $\boldsymbol{S}$ be an unbiased estimator of $\boldsymbol{\Sigma}_*$ under the full model, defined by $\boldsymbol{S} = \boldsymbol{W}_{\omega,0}/(n-k)$, where $\boldsymbol{W}_{\omega,0}$ is the residual matrix in the full model with $\theta = 0$, i.e., $\boldsymbol{W}_{\omega,0} = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_\omega})\boldsymbol{Y}$. By replacing $\boldsymbol{\Sigma}_*$ in (2.3) with $\boldsymbol{S}$, a naive estimator of the risk function can be defined, i.e., the following C_p criterion:

$$C_p(j,\theta) = \mathrm{tr}(\boldsymbol{W}_{j,\theta}\boldsymbol{S}^{-1}) + 2p\mathrm{tr}(\boldsymbol{M}_{j,\theta}^{-1}\boldsymbol{M}_{j,0}). \tag{2.4}$$

However, $C_p(j,\theta)$ has constant bias for $R(j,\theta)$ and it is not negligible when the sample size is small. Hence we try to remove such a bias completely, i.e., our goal is to derive an unbiased estimator of $E^*_{\boldsymbol{Y}}[\mathrm{tr}(\boldsymbol{W}_{j,\theta}\boldsymbol{\Sigma}_*^{-1})]$.

Notice that

$$\boldsymbol{W}_{j,\theta} = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{X}_j\boldsymbol{M}_{j,\theta}^{-1}\boldsymbol{X}_j')\boldsymbol{Y} = \boldsymbol{W}_{\omega,\theta} + \hat{\boldsymbol{\Xi}}'_{\omega,\theta}\boldsymbol{X}'_\omega\boldsymbol{X}_\omega\hat{\boldsymbol{\Xi}}_{\omega,\theta} - \hat{\boldsymbol{\Xi}}'_{j,\theta}\boldsymbol{X}'_j\boldsymbol{X}_j\hat{\boldsymbol{\Xi}}_{j,\theta}.$$

Therefore, it is easy to obtain an unbiased estimator of $E^*_{\boldsymbol{Y}}[\mathrm{tr}(\boldsymbol{W}_{j,\theta}\boldsymbol{\Sigma}_*^{-1})]$ when $\theta = 0$, because $\hat{\boldsymbol{\Xi}}_{j,0}$ and $\boldsymbol{S}$ are independent, and $\hat{\boldsymbol{\Xi}}_{\omega,0}$ and $\boldsymbol{S}$ are also independent. However, when $\theta \neq 0$, it is known that the equation above cannot be used, and that $\hat{\boldsymbol{\Xi}}_{j,\theta}$ and $\boldsymbol{S}$ are not independent, and that $\hat{\boldsymbol{\Xi}}_{\omega,\theta}$ and $\boldsymbol{S}$ are also not independent. Thus, we have to develop an alternative plan to obtain an unbiased estimator of $E^*_{\boldsymbol{Y}}[\mathrm{tr}(\boldsymbol{W}_{j,\theta}\boldsymbol{\Sigma}_*^{-1})]$.

From Draper and Herzberg (1987), we can see that

$$\boldsymbol{W}_{j,\theta} = \boldsymbol{W}_{j,0} + \theta^2\boldsymbol{Y}'\boldsymbol{X}_j\boldsymbol{M}_{j,\theta}^{-1}\boldsymbol{M}_{j,0}\boldsymbol{M}_{j,\theta}^{-1}\boldsymbol{X}_j'\boldsymbol{Y}. \tag{2.5}$$

Notice that $\boldsymbol{W}_{j,0}$ can be rewritten as $\boldsymbol{W}_{\omega,0} + \boldsymbol{Y}'(\boldsymbol{P}_{\boldsymbol{X}_\omega} - \boldsymbol{P}_{\boldsymbol{X}_j})\boldsymbol{Y}$. From this, $\boldsymbol{W}_{j,\theta}$ in (2.5) can also be rewritten as

$$\boldsymbol{W}_{j,\theta} = \boldsymbol{W}_{\omega,0} + \boldsymbol{Y}'(\boldsymbol{P}_{\boldsymbol{X}_\omega} - \boldsymbol{P}_{\boldsymbol{X}_j})\boldsymbol{Y} + \theta^2\boldsymbol{Y}'\boldsymbol{X}_j\boldsymbol{M}_{j,\theta}^{-1}\boldsymbol{M}_{j,0}\boldsymbol{M}_{j,\theta}^{-1}\boldsymbol{X}_j'\boldsymbol{Y}.$$

From this decomposition, it follows that $\boldsymbol{W}_{j,\theta} - \boldsymbol{W}_{\omega,0}$ and $\boldsymbol{S}$ are independent, because $(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_\omega})(\boldsymbol{P}_{\boldsymbol{X}_\omega} - \boldsymbol{P}_{\boldsymbol{X}_j}) = \boldsymbol{O}_n$ and $(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_\omega})\boldsymbol{X}_j\boldsymbol{M}_{j,\theta}^{-1}\boldsymbol{M}_{j,0}\boldsymbol{M}_{j,\theta}^{-1}\boldsymbol{X}_j' = \boldsymbol{O}_n$ are satisfied,

where $\boldsymbol{O}_n$ is the $n \times n$ matrix with all elements zero. Using these independence results, we have

$$\begin{aligned} E_{\boldsymbol{Y}}^{*}\left[\mathrm{tr}(\boldsymbol{W}_{j,\theta}\boldsymbol{S}^{-1})\right] &= (n-k)E_{\boldsymbol{Y}}^{*}\left[\mathrm{tr}\left\{(\boldsymbol{W}_{j,\theta}-\boldsymbol{W}_{\omega,0})\boldsymbol{W}_{\omega,0}^{-1}+\boldsymbol{I}_p\right\}\right] \\ &= (n-k)\left\{E_{\boldsymbol{Y}}^{*}\left[\mathrm{tr}\left\{(\boldsymbol{W}_{j,\theta}-\boldsymbol{W}_{\omega,0})\boldsymbol{\Lambda}\right\}\right]+p\right\} \\ &= (n-k)\left\{E_{\boldsymbol{Y}}^{*}\left[\mathrm{tr}(\boldsymbol{W}_{j,\theta}\boldsymbol{\Lambda})\right]-E_{\boldsymbol{Y}}^{*}\left[\mathrm{tr}(\boldsymbol{W}_{\omega,0}\boldsymbol{\Lambda})\right]+p\right\}, \end{aligned} \tag{2.6}$$

where $\boldsymbol{\Lambda} = E_{\boldsymbol{Y}}^{*}[\boldsymbol{W}_{\omega,0}^{-1}]$. Since $\boldsymbol{W}_{\omega,0} \sim W_p(n-k, \boldsymbol{\Sigma}_*)$, it follows that $E_{\boldsymbol{Y}}^{*}[\boldsymbol{W}_{\omega,0}] = (n-k)\boldsymbol{\Sigma}_*$ and $E_{\boldsymbol{Y}}^{*}[\boldsymbol{W}_{\omega,0}^{-1}] = \boldsymbol{\Sigma}_*^{-1}/(n-k-p-1)$ $(n-k > p+1)$ (see e.g., Siotani, Hayakawa & Fujikoshi, 1985, p. 74, theorem 2.4.6). Substituting the two expectations into (2.6) yields

$$E_{\boldsymbol{Y}}^{*}\left[\mathrm{tr}(\boldsymbol{W}_{j,\theta}\boldsymbol{S}^{-1})\right] = \left(1-\frac{p+1}{n-k}\right)^{-1}\left\{E_{\boldsymbol{Y}}^{*}\left[\mathrm{tr}(\boldsymbol{W}_{j,\theta}\boldsymbol{\Sigma}_*^{-1})\right]-p(p+1)\right\}. \tag{2.7}$$

It follows immediately from the equation (2.7) that an unbiased estimator of $E_{\boldsymbol{Y}}^{*}[\mathrm{tr}(\boldsymbol{W}_{j,\theta}\boldsymbol{\Sigma}_*^{-1})]$ can be defined by $\{1-(p+1)/(n-k)\}\mathrm{tr}(\boldsymbol{W}_{j,\theta}\boldsymbol{S}^{-1})+p(p+1)$. Then, when $n-k > p+1$ holds, we propose the following unbiased estimator of $R(j,\theta)$, which is the modified C_p criterion:

$$MC_p(j,\theta) = \left(1-\frac{p+1}{n-k}\right)\mathrm{tr}(\boldsymbol{W}_{j,\theta}\boldsymbol{S}^{-1}) + 2p\mathrm{tr}(\boldsymbol{M}_{j,\theta}^{-1}\boldsymbol{M}_{j,0}) + p(p+1). \tag{2.8}$$

Notice that $MC_p(j,0)$ coincides with the modified C_p criterion in Fujikoshi and Satoh (1997), which is the information criterion for selecting multivariate linear regression models. Hence, it can be seen that our MC_p is in fact an extended version of Fujikoshi and Satoh's modified C_p.

3. Several Mathematical Properties

In this section, we investigate several mathematical properties of the MC_p and C_p criteria. Let $g(j,\theta)$ be a function of j and θ defined by

$$g(j,\theta) = \mathrm{tr}(\boldsymbol{W}_{j,\theta}\boldsymbol{S}^{-1}). \tag{3.1}$$

By using $g(j,\theta)$ and $C_p(j,\theta)$ in (2.4), $MC_p(j,\theta)$ in (2.8) can be rewritten as

$$MC_p(j,\theta) = C_p(j,\theta) - (1-a)\left\{g(j,\theta)-(n-k)p\right\}, \tag{3.2}$$

where the coefficient a is defined as

$$a = 1-\frac{p+1}{n-k}. \tag{3.3}$$

Notice that the inequality $0 < a < 1$ is satisfied, because $n - k > p + 1$ is true. Thus, the relation $0 < 1 - a < 1$ is also adequate. By substituting this inequality and (A.3) in the Appendix into (3.2), we obtain the following relationship between MC_p and C_p:

THEOREM 1. *For any distribution of* $\boldsymbol{Y}$*, the following inequality is always satisfied:*

$$MC_p(j, \theta) \leq C_p(j, \theta), \tag{3.4}$$

with equality if and only if $\theta = 0$ *and* $j = \omega$*.*

Theorem 1 shows that MC_p is always smaller than C_p, except in the case that the candidate model is the full model with $\theta = 0$. In particular, when the candidate model is the full model and the ridge parameter θ is 0, MC_p and C_p are the same criterion.

Recall that the MC_p criterion is an unbiased estimator of the risk function in (2.3). This unbiasedness, together with Theorem 1, leads to another relation between the MC_p and C_p criteria.

THEOREM 2. *When the distribution of* $\boldsymbol{Y}$ *is normal and the assumption* $\boldsymbol{P}_{\boldsymbol{X}_\omega}\boldsymbol{\Gamma}_* = \boldsymbol{\Gamma}_*$ *is satisfied, the following inequality holds:*

$$E^*_{\boldsymbol{Y}}[MC_p(j, \theta)] = R(j, \theta) \leq E^*_{\boldsymbol{Y}}[C_p(j, \theta)], \tag{3.5}$$

with equality if and only if $\theta = 0$ *and* $j = \omega$*.*

Theorem 2 shows that $MC_p(j, \theta)$ is always an unbiased estimator of $R(j, \theta)$ and $C_p(j, \theta)$ becomes an unbiased estimator of $R(j, \theta)$ when the candidate model is the full model and the ridge parameter θ is 0. Except for the case where the candidate model is the full model with $\theta = 0$, it seems that $C_p(j, \theta)$ overestimates, when compared to $R(j, \theta)$.

Theorem 2 describes biases of the criteria. However, in so far as an information criterion is an estimator of the risk function, not only bias but also variance is an important characteristic, and we now consider the variances of the MC_p and C_p criteria. Let $h(j, \theta)$ be a function of j and θ defined by

$$h(j, \theta) = \mathrm{tr}(\boldsymbol{M}^{-1}_{j,\theta}\boldsymbol{M}_{j,\theta}). \tag{3.6}$$

Using $h(j, \theta)$ and $C_p(j, \theta)$, again we can rewrite $MC_p(j, \theta)$ as

$$MC_p(j, \theta) = aC_p(j, \theta) + 2p(1 - a)h(j, \theta) + p(p + 1), \tag{3.7}$$

where a is given by (3.3). Since p, a and $h(j,\theta)$ are non-stochastic, it seems that variances of MC_p and C_p criteria are related by

$$Var[MC_p(j,\theta)] = a^2 Var[C_p(j,\theta)].$$

Let us recall that $0 < a < 1$. Consequently, we derive the following theorem.

THEOREM 3. *For any distribution of* $\boldsymbol{Y}$*, the following inequality is always satisfied:*

$$Var[MC_p(j,\theta)] < Var[C_p(j,\theta)]. \tag{3.8}$$

Theorem 3 gives us the surprising result that MC_p not only removes the bias of C_p but also reduces the variance. Furthermore, the inequality (3.8) holds even if the distribution of $\boldsymbol{Y}$ is not normal. In general, the variance of a bias-corrected estimator is larger than that of the original estimator (see e.g., Efron & Tibshirani, 1993, p. 138). However, the variance of our MC_p is always smaller than that of C_p, even though MC_p is the bias-corrected C_p. Thus our MC_p criterion has a very desirable property.

Previous theorems have described characteristics of our criterion as an estimator of the risk function. However, in model selection, it is also important which model is chosen by an information criterion. In particular, since we are correcting the bias in the criterion, we need to investigate changes in the selected ridge parameter and/or the selected subset of ω due to this correction of the bias. Let $\hat{\theta}_j^{(m)}$ and $\hat{\theta}_j^{(c)}$ be the ridge parameters minimizing MC_p and C_p criteria respectively, for a fixed j, i.e.,

$$MC_p(j,\hat{\theta}_j^{(m)}) = \min_{\theta\geq 0} MC_p(j,\theta), \quad C_p(j,\hat{\theta}_j^{(c)}) = \min_{\theta\geq 0} C_p(j,\theta). \tag{3.9}$$

Suppose that the inequality $\hat{\theta}_j^{(m)} < \hat{\theta}_j^{(c)}$ holds. Then, from (A.5) in the Appendix, we have $h(j,\hat{\theta}_j^{(c)}) < h(j,\hat{\theta}_j^{(m)})$. Moreover, by applying (3.9) first and then applying (3.4), the relations $MC_p(j,\hat{\theta}_j^{(m)}) \leq MC_p(j,\hat{\theta}_j^{(c)}) \leq C_p(j,\hat{\theta}_j^{(c)})$ can be derived. Substituting the two inequalities into (3.7) yields

$$\begin{aligned} MC_p(j,\hat{\theta}_j^{(m)}) &= aC_p(j,\hat{\theta}_j^{(m)}) + 2p(1-a)h(j,\hat{\theta}_j^{(m)}) + p(p+1) \\ &> aC_p(j,\hat{\theta}_j^{(c)}) + 2p(1-a)h(j,\hat{\theta}_j^{(c)}) + p(p+1) \\ &= MC_p(j,\hat{\theta}_j^{(c)}), \end{aligned}$$

because $0 < a < 1$ and $0 < 1 - a < 1$ are satisfied. However, this result is contradictory to the result that $MC_p(j, \hat{\theta}_j^{(m)}) \leq MC_p(j, \theta)$ for all θ. Consequently, by reductio ad absurdum, we obtain the following theorem which characterizes the relation between two ridge parameters determined by the C_p and MC_p criteria.

THEOREM 4. *For any distribution of* $\boldsymbol{Y}$ *and combinations of* $\boldsymbol{X}$*, the inequality* $\hat{\theta}_j^{(m)} \geq \hat{\theta}_j^{(c)}$ *is always satisfied.*

Theorem 4 shows that the optimal θ obtained using MC_p is not smaller than that determined from C_p for fixed j. In general, the best model obtained from the existing C_p criterion tends to overfit to the principal best model. Many studies have verified this characteristic by conducting numerical simulations, e.g., Fujikoshi and Satoh (1997), Fujikoshi *et al.* (2003), and Fujikoshi, Yanagihara and Wakaki (2005). For the ridge regression model, an overfitting means choosing the smaller θ than the principle best θ $MC_p(j, \theta)$ has been improved so that this weak point is avoided by correcting the bias.

Theorem 4 gives the relation between the two ridge parameters resulting from the MC_p and C_p criteria. By following the approach above that leads to the proof of Theorem 4, we can also obtain other inequalities between the best models resulting from the MC_p and C_p criteria as theorems. (We present the proofs in the Appendix, because they are very similar to the proof of Theorem 4).

THEOREM 5. *Let* $\hat{j}_\theta^{(m)}$ *and* $\hat{j}_\theta^{(c)}$ *be subsets of* ω *minimizing* $MC_p(j, \theta)$ *and* $C_p(j, \theta)$ *respectively, for a fixed* θ*. Then, the relation* $\hat{j}_\theta^{(c)} \not\subset \hat{j}_\theta^{(m)}$ *is always satisfied for any distributions of* $\boldsymbol{Y}$ *and ridge parameters. In particular, for a nested model,* $\hat{j}_\theta^{(m)} \subseteq \hat{j}_\theta^{(c)}$ *holds.*

THEOREM 6. *Let* $\hat{j}^{(m)}$ *and* $\hat{\theta}^{(m)}$ *be* j *and* θ *minimizing* $MC_p(j, \theta)$*, and let* $\hat{j}^{(c)}$ *and* $\hat{\theta}^{(c)}$ *be the values of* j *and* θ *minimizing* $C_p(j, \theta)$*. Then, the inequality* $\hat{\theta}^{(m)} \geq \hat{\theta}^{(c)}$ *or the relation* $\hat{j}_\theta^{(c)} \not\subset \hat{j}_\theta^{(m)}$ *are always satisfied for any distributions of* $\boldsymbol{Y}$*.*

4. Numerical Study

We evaluate the proposed criterion applied numerically to the polynomial regression model, $\boldsymbol{Y} \sim N_{n \times p}(\boldsymbol{\Gamma}_*, \boldsymbol{\Sigma}_* \otimes \boldsymbol{I}_p)$, with $p = 2$, $n = 20$, $k = 12$ and $\omega = \{1, \cdots, 12\}$ where

$\boldsymbol{\Gamma}_* = \boldsymbol{X}_\omega \boldsymbol{\Xi}_*$,

$$\boldsymbol{\Xi}_* = \delta \begin{pmatrix} 1 & 2 & 3 & 0 & \cdots & 0 \\ 1 & 4 & 9 & 0 & \cdots & 0 \end{pmatrix}', \ \boldsymbol{\Sigma}_* = \begin{pmatrix} 1 & 0.5^2 \\ 0.5^2 & 1 \end{pmatrix},$$

$$\boldsymbol{X}_\omega = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{pmatrix} \text{ and } \boldsymbol{Z} = \begin{pmatrix} z_1 & z_1^2 & \cdots & z_1^k \\ \vdots & \vdots & \vdots & \vdots \\ z_n & z_n^2 & \cdots & z_n^k \end{pmatrix}.$$

Each column vector of the design matrix $\boldsymbol{X}_\omega$ is given by standardization of the corresponding column vector of $\boldsymbol{Z}$. The first column vector of $\boldsymbol{Z}$ is generated from the independent uniform distribution on $(-1, 1)$. Note that the candidate models are nested and $\boldsymbol{X}_j$ is the submatrix consisting of the first j columns of $\boldsymbol{X}_\omega$. In a sense, the subindex j is the degree of a polynomial here.

Since by our criterion, MC_p is derived as an estimator of the MSE of prediction, we compare the related four criteria: MC_p, C_p, the cross-validation (CV) criterion (Stone, 1974) and the generalized cross validation (GCV) criterion (Craven & Wahba, 1979), on the following three points, (*i*) the probabilities or frequencies of selected models, (*ii*) the expectation value of the selected ridge parameter, (*iii*) the MSE of prediction. Here, CV and GCV criteria can be formally defined by

$$\begin{aligned} \mathrm{CV}(j, \theta) &= \sum_{i=1}^{n} \frac{((\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{j,\theta})\boldsymbol{S}^{-1}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{j,\theta})')_{ii}}{\{1 - (\boldsymbol{X}_j \boldsymbol{M}_{j,\theta}^{-1} \boldsymbol{X}_j')_{ii}\}^2}, \\ \mathrm{GCV}(j, \theta) &= \frac{\mathrm{tr}(\boldsymbol{W}_{j,\theta}\boldsymbol{S}^{-1})}{\{1 - \mathrm{tr}(\boldsymbol{M}_{j,\theta}^{-1}\boldsymbol{M}_{j,0})/n\}^2}, \end{aligned} \tag{4.1}$$

where $(\boldsymbol{A})_{ii}$ denotes the (i, i)th element of a matrix $\boldsymbol{A}$. We selected both the candidate model and the ridge parameter, and the MSE of the prediction as $np + E_{\boldsymbol{Y}}^*[\mathrm{tr}\{(\boldsymbol{\Gamma}_* - \hat{\boldsymbol{Y}}_{\hat{j},\hat{\theta}})\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\Gamma}_* - \hat{\boldsymbol{Y}}_{\hat{j},\hat{\theta}})'\}]$. Those properties were evaluated by Monte Carlo simulation with 1,000 iterations under two types of true model, 1) $\delta = 0$, or a constant model, 2) $\delta = 2.0$, or a third degree polynomial model. In the former case, smaller degree polynomial models estimated by larger ridge parameters should be selected, conversely, the third degree polynomial model estimated by smaller ridge parameters should be selected in the latter case.

As the result of the simulation study, our MC_p criterion was much improved, compared to the original Mallows' C_p criterion in the sense of the MSE of prediction. Although the MSE was almost the same for the MC_p and CV criteria, MC_p selected preferable candidate models more often than CV in both of the cases 1) when larger ridge parameters

were required, 2) when ridge parameters were not as necessary, or the usual least square estimator without ridge parameters was sufficient. The performance of the GCV criterion might be located in between that of the CV and C_p criteria. Therefore we conclude that the MC_p criterion is the best criterion among those four criteria in the sense of MSE prediction and the probability of selecting preferable candidate models for a ridge regression model.

5. Conclusion and Discussion

In this paper, we have proposed an unbiased C_p criterion, denoted as MC_p. The MC_p criterion is an unbiased estimator of the risk function based on the expected standardized MSE of prediction when the distribution of $\boldsymbol{Y}$ is normal and $\boldsymbol{P}_{\boldsymbol{X}_\omega}\boldsymbol{\Gamma}_* = \boldsymbol{\Gamma}_*$ is satisfied. One of advantages of the MC_p criterion is that its definition is very simple. Furthermore, we have proved analytically that the MC_p criterion has smaller variance than the C_p criterion. In addition, the optimal ridge parameter obtained using MC_p is always at least as large as that resulting from C_p for fixed j, and the best subset of ω obtained by using MC_p is not included in the best subset obtained from use of C_p. In numerical studies, we demonstrated that the MC_p criterion is more effective than the C_p criterion and the formal CV and GCV criteria.

Davies, Neath and Cavanaugh (2006) showed that $MC_p(j, 0)$ is the minimum unbiased estimator of $R(j, 0)$ when the candidate model is the true model in the case of $p = 1$. Moreover, from the asymptotic expansion of the bias in Fujikoshi, Yanagihara and Wakaki (2005), it seems that the effect of non-normality on the bias of $MC_p(j, 0)$ is very small; its order is merely $O(n^{-2})$ even under non-normality, when $\boldsymbol{P}_{\boldsymbol{X}_\omega}\boldsymbol{\Gamma}_* = \boldsymbol{\Gamma}_*$ is satisfied. Hence, we can expect that our $MC_p(j, \theta)$ also has similar good properties.

When the observations are univariate, the risk function does not need to be standardized by the true variance-covariance matrix and in this case, an unbiased estimator is easy to obtain. This unbiased estimator may almost be equivalent to the criterion proposed by Li (1986). However, when observations are multivariate, the standardization results in difficulty deriving an unbiased estimator. For this case, there has been no unbiased estimator of the risk based on the expected standardized MSE of prediction for multivariate ridge regression models. On the other hand, we can guess that an estimator of the risk function may be able to be derived easily by use of the CV method as in (4.1),

even when the observations are multivariate. However, in the multivariate case, an estimated variance-covariance matrix for the standardization should also be constructed by the jackknife method, as well as by using the predictor of $\boldsymbol{Y}$. Then, the GCV criterion cannot be strictly defined and the CV criterion will have constant bias (see Fujikoshi *et al.*, 2003). Therefore, for the selection of a multivariate ridge regression model, MC_p will not be supplanted by the other criteria at present.

There have been many studies concerned with correction of the bias of an information criterion. However, in almost all cases the resulting papers have reported only on the bias correction and have not discussed the difference in variable selection, by comparison of the original criterion and the theoretically improved version. In contrast, this paper does consider changes in the selected model due to correcting the bias.

From the many viewpoints mentioned above, we consider that the results in our paper are useful, and thus we can recommend use of the MC_p criterion instead of the C_p criterion for model selection for multivariate ridge regression models.

Appendix

A.1. Properties of the function $g(j, \theta)$

Let $\boldsymbol{P}_j$ be a $k_j \times k_j$ orthogonal matrix such that

$$\boldsymbol{P}_j' \boldsymbol{M}_{j,0} \boldsymbol{P}_j = \boldsymbol{D}_j = \mathrm{diag}(d_{j,1}, \ldots, d_{j,k_j}), \tag{A.1}$$

where $d_{j,\alpha}$ $(\alpha = 1, \ldots, k_j)$ are eigenvalues of $\boldsymbol{M}_{j,0}$, and let $\boldsymbol{z}_{j,1}, \ldots, \boldsymbol{z}_{j,k_j}$ be $n \times 1$ vectors such that $(\boldsymbol{z}_{j,1}, \ldots, \boldsymbol{z}_{j,k_j})' = \boldsymbol{P}_j' \boldsymbol{X}_j' \boldsymbol{Y} \boldsymbol{S}^{-1/2}$. By using $\boldsymbol{z}_{j,\alpha}$ and $d_{j,\alpha}$ $(\alpha = 1, \ldots, k_j)$, we can write $g(j, \theta)$ in (3.1) as

$$g(j, \theta) = \mathrm{tr}(\boldsymbol{Y}'\boldsymbol{Y}\boldsymbol{S}^{-1}) - 2 \sum_{\alpha=1}^{k_j} \frac{\|\boldsymbol{z}_{j,\alpha}\|^2}{d_{j,\alpha} + \theta} + \sum_{\alpha=1}^{k_j} \frac{\|\boldsymbol{z}_{j,\alpha}\|^2 d_{j,\alpha}}{(d_{j,\alpha} + \theta)^2}. \tag{A.2}$$

Since $d_{j,\alpha} > 0$ and $\theta \geq 0$ hold, we have

$$\frac{\partial}{\partial \theta} g(j, \theta) = 2\theta \sum_{\alpha=1}^{k_j} \frac{\|\boldsymbol{z}_{j,\alpha}\|^2}{(d_{j,\alpha} + \theta)^3} \geq 0,$$

with equality if and only if $\theta = 0$ or $\theta \to \infty$. Therefore, we can see that $g(j, \theta)$ is a strictly monotonic increasing function of $\theta \in [0, \infty]$. This result implies that $g(j, \theta) \geq g(j, 0)$ with

equality if and only if $\theta = 0$. Notice that $\boldsymbol{P}_{\boldsymbol{X}_\omega} - \boldsymbol{P}_{\boldsymbol{X}_j}$ is positive definite except when $j = \omega$. Therefore, we obtain

$$g(j,0) = \mathrm{tr}(\boldsymbol{W}_{j,0}\boldsymbol{S}^{-1}) = (n-k)p + \mathrm{tr}\{\boldsymbol{Y}'(\boldsymbol{P}_{\boldsymbol{X}_\omega} - \boldsymbol{P}_{\boldsymbol{X}_j})\boldsymbol{Y}\boldsymbol{S}^{-1}\} \geq (n-k)p = g(\omega,0),$$

with equality if and only if $j = \omega$. Results obtained in this subsection are summarized in the following theorem.

THEOREM A.1. *The function $g(j,\theta)$ is a strictly monotonic increasing function of $\theta \in [0,\infty]$ for fixed j, and has the following lower bound:*

$$g(j,\theta) \geq (n-k)p, \tag{A.3}$$

with equality if and only if $\theta = 0$ and $j = \omega$.

A.2. Monotonicity of the function $h(j,\theta)$

Notice that the function $h(j,\theta)$ in (3.6) can be rewritten as

$$h(j,\theta) = \sum_{\alpha=1}^{k_j} \frac{d_{j,\alpha}}{d_{j,\alpha}+\theta}, \tag{A.4}$$

where $d_{j,\alpha}$ $(\alpha = 1,\ldots,k_j)$ are the eigenvalues of $\boldsymbol{M}_{j,0}$, which are defined by (A.1). From the equation (A.4), we have the following theorem.

THEOREM A.2. *The function $h(j,\theta)$ is a strictly decreasing function of $\theta \in [0,\infty]$ for fixed j. Therefore, we have the following relation:*

$$h(j,\theta_2) < h(j,\theta_1), \quad (\textit{when } \theta_1 < \theta_2). \tag{A.5}$$

Let $\boldsymbol{X}_{j_1} = (\boldsymbol{X}_j\ \boldsymbol{x})$ be an $n \times (k_j+1)$ matrix, where $\boldsymbol{x}$ is an $n \times 1$ vector which is linearly independent of any row vectors of $\boldsymbol{X}_j$. From the formula for the inverse matrix (see e.g., Siotani, Hayakawa & Fujikoshi, 1985, p. 592, theorem A.2.3), we have

$$\boldsymbol{M}_{j_1,\theta}^{-1} = \begin{pmatrix} \boldsymbol{M}_{j,\theta}^{-1} - \boldsymbol{M}_{j,\theta}^{-1}\boldsymbol{X}_j'\boldsymbol{x}\boldsymbol{x}'\boldsymbol{X}_j\boldsymbol{M}_{j,\theta}^{-1}/c_{j,\theta} & -\boldsymbol{M}_{j,\theta}^{-1}\boldsymbol{X}_j'\boldsymbol{x}/c_{j,\theta} \\ -\boldsymbol{x}'\boldsymbol{X}_j\boldsymbol{M}_{j,\theta}^{-1}/c_{j,\theta} & 1/c_{j,\theta} \end{pmatrix}, \tag{A.6}$$

where $c_{j,\theta} = \theta + \boldsymbol{x}'(\boldsymbol{I}_n - \boldsymbol{X}_j\boldsymbol{M}_{j,\theta}^{-1}\boldsymbol{X}_j')\boldsymbol{x}$. Let $b_{j,1},\ldots,b_{j,k_j}$ be such that $(b_{j,1},\ldots,b_{j,k_j})' = \boldsymbol{P}_j'\boldsymbol{X}_j'\boldsymbol{x}$. By using $b_{j,\alpha}$ and $d_{j,\alpha}$ $(\alpha = 1,\ldots,k_j)$, we can write $c_{j,\theta}$ as

$$c_{j,\theta} = \theta + \boldsymbol{x}'\boldsymbol{x} - \sum_{\alpha=1}^{k_j} \frac{b_{j,\alpha}^2}{d_{j,\alpha}+\theta}. \tag{A.7}$$

By partially differentiating the equation (A.7), we can see that $c_{j,\theta}$ is a monotonic decreasing function of θ. This implies that $c_{j,\theta} \geq c_{j,0} = \boldsymbol{x}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_j})\boldsymbol{x}$. Notice that $\boldsymbol{x}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_j})\boldsymbol{x} > 0$ because $\boldsymbol{P}_{\boldsymbol{X}_j}\boldsymbol{x} \neq \boldsymbol{x}$. Hence it follows that $c_{j,\theta} > 0$. Moreover, from (A.6), $h(j_1, \theta)$ is given by

$$h(j_1, \theta) = h(j, \theta) + \frac{1}{c_{j,\theta}} \boldsymbol{x}'(\boldsymbol{I}_n - \boldsymbol{X}_j \boldsymbol{M}_{j,\theta} \boldsymbol{X}_j')^2 \boldsymbol{x}. \tag{A.8}$$

By applying a similar expression in (A.7), we have

$$\boldsymbol{x}'(\boldsymbol{I}_n - \boldsymbol{X}_j \boldsymbol{M}_{j,\theta} \boldsymbol{X}_j')^2 \boldsymbol{x} = \boldsymbol{x}'\boldsymbol{x} - 2 \sum_{\alpha=1}^{k_j} \frac{b_{j,\alpha}^2}{d_{j,\alpha} + \theta} + \sum_{\alpha=1}^{k_j} \frac{b_{j,\alpha}^2 d_{j,\alpha}}{(d_{j,\alpha} + \theta)^2}.$$

The above equation leads to

$$\frac{\partial}{\partial \theta} \boldsymbol{x}'(\boldsymbol{I}_n - \boldsymbol{X}_j \boldsymbol{M}_{j,\theta} \boldsymbol{X}_j')^2 \boldsymbol{x} = 2\theta \sum_{\alpha=1}^{k_j} \frac{b_{j,\alpha}^2}{(d_{j,\alpha} + \theta)^3} \geq 0,$$

with equality if and only if $\theta = 0$ or $\theta \to \infty$. Therefore, we can see that $\boldsymbol{x}'(\boldsymbol{I}_n - \boldsymbol{X}_j \boldsymbol{M}_{j,\theta} \boldsymbol{X}_j')^2 \boldsymbol{x}$ is a strictly monotonic increasing function of $\theta \in [0, \infty]$. From this result, we obtain

$$\boldsymbol{x}'(\boldsymbol{I}_n - \boldsymbol{X}_j \boldsymbol{M}_{j,\theta} \boldsymbol{X}_j')^2 \boldsymbol{x} \geq \boldsymbol{x}'(\boldsymbol{I}_n - \boldsymbol{X}_j \boldsymbol{M}_{j,0} \boldsymbol{X}_j')^2 \boldsymbol{x} = \boldsymbol{x}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_j})\boldsymbol{x} > 0.$$

Notice that $c_{j,\theta} > 0$. Substituting the above inequality into (A.8) yields $h(j_1, \theta) > h(j, \theta)$ when $\boldsymbol{X}_{j_1} = (\boldsymbol{X}_j \ \boldsymbol{x})$. By means of similar calculations, we obtain the following theorem.

THEOREM A.3. *For fixed θ, the following relation on $h(j, \theta)$ can be derived:*

$$h(j_1, \theta) < h(j_2, \theta), \quad (\textit{when } j_1 \subset j_2). \tag{A.9}$$

A.3. Proofs of Theorems 5 and 6

First, we give the proof of Theorem 5. Recall that $\hat{j}_\theta^{(m)}$ and $\hat{j}_\theta^{(c)}$ are j minimizing $MC_p(j, \theta)$ and $C_p(j, \theta)$ respectively, for fixed θ. Then, we have

$$MC_p(\hat{j}_\theta^{(m)}, \theta) = \min_{j \subseteq \omega} MC_p(j, \theta), \quad C_p(\hat{j}_\theta^{(c)}, \theta) = \min_{j \subseteq \omega} C_p(j, \theta). \tag{A.10}$$

Suppose that the inequality $\hat{j}_\theta^{(c)} \subset \hat{j}_\theta^{(m)}$ holds. Then, from (A.9) in the Appendix, we derive $h(\hat{j}_\theta^{(m)}, \theta) < h(\hat{j}_\theta^{(c)}, \theta)$. Moreover, by applying (A.10) first and then applying (3.4),

$MC_p(\hat{j}_\theta^{(m)}, \theta) \leq MC_p(\hat{j}_\theta^{(c)}, \theta) \leq C_p(\hat{j}_\theta^{(c)}, \theta)$ are obtained. Notice that $0 < a < 1$ and $0 < 1 - a < 1$. Substituting the two inequalities into (3.7) yields

$$\begin{aligned} MC_p(\hat{j}_\theta^{(m)}, \theta) &= aC_p(\hat{j}_\theta^{(m)}, \theta) + 2p(1-a)h(\hat{j}_\theta^{(m)}, \theta) + p(p+1) \\ &> aC_p(\hat{j}_\theta^{(c)}, \theta) + 2p(1-a)h(\hat{j}_\theta^{(c)}, \theta) + p(p+1) \\ &= MC_p(\hat{j}_\theta^{(c)}, \theta). \end{aligned}$$

However, this result is contradictory to $MC_p(\hat{j}_\theta^{(m)}, \theta) \leq MC_p(j, \theta)$ for all j. Consequently, by reductio ad absurdum, the statement in Theorem 5 is correct.

Next, we give the proof of Theorem 6. Let us recall that $\hat{\theta}^{(m)}$ and $\hat{j}^{(m)}$ are θ and j minimizing $MC_p(j, \theta)$, and let $\hat{\theta}^{(c)}$, $\hat{j}^{(c)}$ be θ, j minimizing $C_p(j, \theta)$. Then, we have

$$MC_p(\hat{j}^{(m)}, \hat{\theta}^{(m)}) = \min_{j \subseteq \omega, \theta \geq 0} MC_p(j, \theta), \quad C_p(\hat{j}^{(c)}, \hat{\theta}_j^{(c)}) = \min_{j \subseteq \omega, \theta \geq 0} C_p(j, \theta). \tag{A.11}$$

Suppose that the inequalities $\hat{\theta}^{(m)} < \hat{\theta}^{(c)}$ and $\hat{j}^{(c)} \subset \hat{j}^{(m)}$ hold. Then, from (A.5) and (A.9) in the Appendix, we have $h(\hat{j}^{(c)}, \hat{\theta}^{(c)}) < h(\hat{j}^{(m)}, \hat{\theta}^{(m)})$. Moreover, by applying (A.10) first and then applying (3.4), the inequalities $MC_p(\hat{j}^{(m)}, \hat{\theta}^{(m)}) \leq MC_p(\hat{j}^{(c)}, \hat{\theta}^{(c)}) \leq C_p(\hat{j}^{(c)}, \hat{\theta}^{(c)})$ are obtained. Notice that $0 < a < 1$ and $0 < 1 - a < 1$. Substituting the two inequalities into (3.7) yields

$$\begin{aligned} MC_p(\hat{j}^{(m)}, \hat{\theta}^{(m)}) &= aC_p(\hat{j}^{(m)}, \hat{\theta}^{(m)}) + 2p(1-a)h(\hat{j}^{(m)}, \hat{\theta}^{(m)}) + p(p+1) \\ &> aC_p(\hat{j}^{(c)}, \hat{\theta}^{(c)}) + 2p(1-a)h(\hat{j}^{(c)}, \hat{\theta}^{(c)}) + p(p+1) \\ &= MC_p(\hat{j}^{(c)}, \hat{\theta}^{(c)}). \end{aligned}$$

However, this result is contradictory to $MC_p(\hat{j}^{(m)}, \hat{\theta}^{(m)}) \leq MC_p(j, \theta)$ for all θ and j. Consequently, by reductio ad absurdum, it follows that the statement of Theorem 6 is correct.

Acknowledgment

A part of this research was supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), #19700265, 2007–2010.

References

[1] Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 377–403.

[2] Davies, S. L., Neath, A. A. & Cavanaugh, J. E. (2006). Estimation optimality of corrected AIC and modified C_p in linear regression model. *International Statist. Review*, **74**, 161–168.

[3] Draper, N. R. & Herzberg, A. M. (1987). A ridge-regression sidelight. *Amer. Statist.*, **41**, 282–283.

[4] Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CLC, New York.

[5] Fujikoshi, Y. & Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika*, **84**, 707–716.

[6] Fujikoshi, Y., Yanagihara, H. & Wakaki, H. (2005). Bias corrections of some criteria for selection multivariate linear regression models in a general case. *Amer. J. Math. Management Sci.*, **25**, 221–258.

[7] Fujikoshi, Y., Noguchi, T., Ohtaki, M. & Yanagihara, H. (2003). Corrected versions of cross-validation criteria for selecting multivariate regression and growth curve models. *Ann. Inst. Statist. Math.*, **55**, 537–553.

[8] Hurvich, C. M., Simonoff, J. S. & Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. Roy. Statist. Soc. Ser.* **B**, **60**, 271–293.

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

[10] Mallows, C. L. (1995). More comments on C_p. *Technometrics*, **37**, 362–372.

[11] Li, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.*, **14**, 1101–1112.

[12] Siotani, M., Hayakawa, T. & Fujikoshi, Y. (1985). *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook*. American Sciences Press, Columbus, Ohio.

[13] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser.* **B**, **36**, 111–147.

TABLE 1. The frequencies of selected models, the expectation value of selected ridge parameters, the MSE of prediction for 1,000 repetitions under the true model with $\delta = 0$, or a constant model.

	MC_p		C_p		CV		GCV	
j	freq.	$E[\hat{\theta}_j]$	freq.	$E[\hat{\theta}_j]$	freq.	$E[\hat{\theta}_j]$	freq.	$E[\hat{\theta}_j]$
1	639	82.9	430	70.7	513	75.7	496	76.4
2	130	83.7	140	68.0	152	74.5	152	75.3
3	25	84.1	38	67.1	48	74.9	42	76.8
4	24	84.9	26	66.6	40	76.2	37	77.6
5	13	84.7	24	66.4	29	77.4	20	77.6
6	17	85.0	34	66.6	24	78.3	27	78.0
7	15	85.0	32	65.8	14	77.9	21	77.7
8	14	85.6	36	65.0	11	78.6	26	78.2
9	12	85.4	22	64.6	15	78.3	16	78.3
10	14	85.3	39	64.1	21	78.4	23	78.6
11	21	85.1	46	64.4	23	78.3	30	77.9
12	76	85.3	133	65.0	110	78.0	110	78.1
MSE	42.9		46.6		42.9		44.1	

TABLE 2. The frequencies of selected models, the expectation value of selected ridge parameters, the MSE of prediction for 1,000 repetitions under the true model with $\delta = 2.0$, or a third degree polynomial model.

	MC_p		C_p		CV		GCV	
j	freq.	$E[\hat{\theta}_j]$	freq.	$E[\hat{\theta}_j]$	freq.	$E[\hat{\theta}_j]$	freq.	$E[\hat{\theta}_j]$
1	0	0.01	0	0.01	0	0.00	0	0.15
2	0	0.02	0	0.01	0	0.00	0	0.21
3	757	0.01	520	0.01	614	0.02	612	0.01
4	80	0.04	98	0.02	153	0.03	109	0.03
5	36	0.06	56	0.03	85	0.08	58	0.03
6	18	0.07	48	0.03	50	0.12	45	0.04
7	15	0.05	32	0.02	27	0.09	25	0.03
8	15	0.06	31	0.03	14	0.13	24	0.03
9	9	0.06	28	0.03	10	0.12	18	0.03
10	13	0.06	38	0.03	6	0.15	24	0.03
11	23	0.07	57	0.03	12	0.16	36	0.04
12	34	0.07	92	0.03	29	0.17	49	0.03
MSE	48.8		52.0		48.9		50.3	