Prediction Error Criterion for Selecting of Variables in Regression Model

Yasunori Fujikoshi*, Tamio, Kan**
and Shin Takahashi**

*Department of Mathematics, Graduate School of Science and Engineering, Chuo University, Bunkyo-ku, Tokyo, Japan
and **Esumi Co., Ltd., Tokyo, Japan

Abstract

This paper is concerned with criteria for selecting of variables in regression model. We propose a prediction error criterion C_{pe} which is unbiased as an estimator for the mean squared error in prediction R_{pe}, when the true model is contained in the full model. The property is shown without normality. Such unbiasedness property is studied for other criteria such as cross-validation criterion, C_p criterion, etc. We will also examine modifications of multiple correlation coefficient from the point of estimating R_{pe}. Our results are extended to multivariate case.

AMS 2000 Subject Classification: primary 62H10; secondary 62E20

Key Words and Phrases: Prediction error criterion, Regression models, Selection of variables, Unbiasedness.

Abbreviated title: Prediction Error Criterion in Regression Model

1. Introduction

In univariate regression model, we want to predict or describe a response variable y by a subset of several explanatory variables $x_1, \ldots, x_k$. Suppose that there are n observations on y and $\boldsymbol{x} = (x_1, \ldots, x_k)'$ denoted by $y_\alpha, \boldsymbol{x}_\alpha = (x_{\alpha 1}, \ldots, x_{\alpha k})'; \alpha = 1, \ldots, n$. In this paper we consider the problem of selecting a model from a collection of candidate models specified by linear regression of y on subvectors of $\boldsymbol{x}$. We assume that the true model for $y_\alpha, \alpha = 1, \ldots, n$ is as follows:

$$M_0 : \; y_\alpha = \eta_{\alpha 0} + \varepsilon_{\alpha 0}, \; \alpha = 1, \ldots, n, \tag{1.1}$$

where the error terms $\varepsilon_{10}, \ldots, \varepsilon_{n0}$ are mutually independent, and each of them has the same mean 0 and the same variance σ_0^2. The linear regression model including all the explanatory variables is written as

$$M_F : \; y_\alpha = \beta_0 + \beta_1 x_{\alpha 1} + \ldots + \beta_k x_{\alpha k} + \varepsilon_\alpha, \; \alpha = 1, \ldots, n, \tag{1.2}$$

where the coefficients $\beta_0, \ldots, \beta_j$ are unknown parameters, the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are matually independent and have the same mean 0 and the same unknown variance σ^2. The model is called the full model.

In this paper we are interested in criteria for selecting of models, more concretely for selecting of variables. As a subset of all explanatory variables, without loss of generality we may consider the subset of the first j explanatory variables $x_1, \ldots, x_j$. Consider a candidate model

$$M_J : \; y_\alpha = \beta_0 + \beta_1 x_{\alpha 1} + \ldots + \beta_j x_{\alpha j} + \varepsilon_\alpha, \; \alpha = 1, \ldots, n, \tag{1.3}$$

where the coefficient $\beta_o, \ldots, \beta_j$ are unkown, and the error terms are the same ones as in (1.2).

As a criterion for goodness of a fitted candidate model we consider the prediction errors, more precisely the mean squares errors in prediction. The measure is given by

$$R_{pe} = \sum_{\alpha=1}^{n} \mathrm{E}_0[(z_\alpha - \hat{y}_{\alpha J})^2], \tag{1.4}$$

where $\hat{y}_{\alpha J}$ is the usual unbiased estimator of $\eta_{\alpha 0}$ under M_J, and $\boldsymbol{z} = (z_1, \ldots, z_n)'$ has the same distribution as $\boldsymbol{y}$ in (1.1) and is independent of $\boldsymbol{y}$. We call R_{pe} a risk function for M_J. Here E_0 denotes the expectation with respect to the true model M_0. It is easy to see that

$$R_{pe} = \sum_{\alpha=1}^{n} \mathrm{E}_0[(\eta_{\alpha 0} - \hat{y}_{\alpha J})^2] + n\sigma_0^2. \tag{1.5}$$

Therefore, the target criterion is essentially the same as the first term of the right-hand side in (1.5). A typical estimation method for (1.4) is to use a cross-validiation method (see, e.g., Sone (1974)). The method predicts y_α by the usual unbiased estimator $\hat{y}_{(-\alpha)J}$ based on the data set obtained by removing the α-th observation $(y_\alpha, \boldsymbol{x}'_\alpha)$, and estimates R_{pe} by

$$C_{cv} = \sum_{\alpha=1}^{n} \{y_\alpha - \hat{y}_{(-\alpha)J}\}^2. \tag{1.6}$$

The selection method is to choose the model for which C_{cv} is minimized. If the errors are normally distributed, we can use a well known AIC (Akakike (1973)) which are not discussed here.

In this paper we propose a new criterion

$$C_{pe} = s_J^2 + \frac{2(j+1)}{n-k-1} s_F^2, \tag{1.7}$$

where s_J^2 and s_F^2 are is the sums of squares of residuals in the candidate model M_J and the full model M_F, respectively.

In Section 2 we study unbiasedness properties of C_{cv} and C_{pe} as an estimator for their target measure R_{pe}. It is shown that C_{cv} is only asymptotically unbiased while C_{pe} is exactly unbiased when the true model is contained in the full model. In Section 3 we shall make clear a relationship of C_{pe} with C_p (Mallows (1973)) and its modification C_{mp} (Fujikoshi and Satoh (1997)). The latter criteria are cosely related to C_{pe}, since the target mesure for C_p and C_{mp} is R_{pe}/σ_0^2. In Section 4 we also propose an adjusted multiple correlation coefficient and its monotone transformation given by

$$\bar{R}^2 = 1 - \frac{n+j+1}{n-j-1}(1-R^2),$$
$$\bar{C}_{dc} = (1-\bar{R}^2)s_y^2 = \frac{n+j+1}{n-j-1} s_J^2,$$

where R is the multiple correlation coefficient between y and $(x_1, \ldots, x_j)$, and $s_y^2/(n-1)$ is the usual sample variance of y. We show that $\bar{C}_{dc}$ is an unbiased estimator of R_{pe} when the true model is contained in the model M_J. In Section 5 we give a multivariate extension of C_{pe}. A numerical example is given in Section 6.

2 Unbiasedness of C_{cv} and C_{pe}

A naive estimator for R_{pe} is obtained by substituting y_α to z_α in (1.4), therefore by yielding

$$\sum_{\alpha=1}^{n}(y_\alpha - \hat{y}_{\alpha J})^2 = s_J^2.$$

Writing the model M_J as in matrix form, we have

$$\boldsymbol{y} = (y_1, \ldots, y_n)' = X_J\boldsymbol{\beta}_J + (\varepsilon_1, \ldots, \varepsilon_n)',$$

where $\boldsymbol{\beta}_J = (\beta_0, \beta_1, \ldots, \beta_j)'$, and X_J is the matrix constructed from the first $j+1$ columns of $X = (\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n)'$ with $\tilde{\boldsymbol{x}}_\alpha = (1\ \boldsymbol{x}'_\alpha)'$. The best linear predictor under the model M_J is expressed as

$$\begin{aligned}\hat{\boldsymbol{y}}_J &= (\hat{y}_{1J}, \ldots, \hat{y}_{nJ})' \\ &= X_J(X'_JX_J)^{-1}X'_J\boldsymbol{y} = P_J\boldsymbol{y},\end{aligned}$$

where $P_J = X_J(X'_JX_J)^{-1}X'_J$ is a projection matrix of the space $\mathcal{R}[X_J]$ spanned by the column vectors of X_J

Lemma 2.1 *The risk R_{pe} for the model M_J in* (1.4) *is written as*

$$R_{pe} = \mathrm{E}_0(s_J^2) + B_{pe}, \tag{2.1}$$

where

$$B_{pe} = 2(j+1)\sigma_0^2. \tag{2.2}$$

Further,

$$\mathrm{E}_0(s_J^2) = (n-j-1)\sigma_0^2 + \delta_J^2, \tag{2.3}$$

where $\delta_J^2 = \boldsymbol{\eta}'_0(I_n - P_J)\boldsymbol{\eta}_0$, and if the true model is contained in the model M_J,

$$R_{pe} = (n+j+1)\sigma_0^2. \tag{2.4}$$

Proof

Note that

$$B_{pe} = \mathrm{E}_0[(\boldsymbol{z} - \hat{\boldsymbol{y}}_J)'(\boldsymbol{z} - \hat{\boldsymbol{y}}_J) - (\boldsymbol{y} - \hat{\boldsymbol{y}}_J)'(\boldsymbol{y} - \hat{\boldsymbol{y}}_J)].$$

We have

$$
\begin{aligned}
&\mathrm{E}_0\left[(\boldsymbol{z}-\hat{\boldsymbol{y}}_J)'(\boldsymbol{z}-\hat{\boldsymbol{y}}_J)\right] \\
&\quad = \mathrm{E}_0\left[\{\boldsymbol{z}-\boldsymbol{\eta}_0 - P_J(\boldsymbol{y}-\boldsymbol{\eta}_0) + (I-P_J)\boldsymbol{\eta}_0\}'\right. \\
&\qquad \left.\times \{\boldsymbol{z}-\boldsymbol{\eta}_0 - P_J(\boldsymbol{y}-\boldsymbol{\eta}_0) + (I-P_J)\boldsymbol{\eta}_0\}\right] \\
&\quad = n\sigma_0^2 + (j+1)\sigma_0^2 + \delta_J^2, \\
&\mathrm{E}_0\left[(\boldsymbol{y}-\hat{\boldsymbol{y}}_J)'(\boldsymbol{y}-\hat{\boldsymbol{y}}_J)\right] \\
&\quad = \mathrm{E}_0\left[\{\boldsymbol{y}-\boldsymbol{\eta}_0 - P_J(\boldsymbol{y}-\boldsymbol{\eta}_0) + (I-P_J)\boldsymbol{\eta}_0)\}'\right. \\
&\qquad \left.\times \{\boldsymbol{y}-\boldsymbol{\eta}_0 - P_J(\boldsymbol{y}-\boldsymbol{\eta}_0) + (I-P_J)\boldsymbol{\eta}_0\}\right] \\
&\quad = n\sigma_0^2 - (j+1)\sigma_0^2 + \delta_J^2.
\end{aligned}
$$

Note that $s_J^2 = (\boldsymbol{y}-\hat{\boldsymbol{y}})'(\boldsymbol{y}-\hat{\boldsymbol{y}})$. Therefore, from the above results our conclusions are obtained.

Theorem 2.1 *Suppose that the true model M_0 is contained in the full model M_F. Then, the criterion C_{pe} defined by* (1.7) *is an exact unbiased estimator for R_{pe}.*

Proof

From Lemma 2.1 we have

$$R_{pe} = \mathrm{E}_0(s_J^2) + 2(j+1)\sigma_0^2.$$

Note that $s_F^2 = \boldsymbol{y}'(I_n - P_F)\boldsymbol{y}$, where $P_F = X(X'X)^{-1}X'$. Since the true model M_0 is contained in the full model M_F, $P_F\boldsymbol{\eta}_0 = \boldsymbol{\eta}_0$, and we have

$$
\begin{aligned}
\mathrm{E}(s_F^2) &= \mathrm{E}[(\boldsymbol{y}-\boldsymbol{\eta}_0)'(I_n - P_F)(\boldsymbol{y}-\boldsymbol{\eta}_0)] \\
&= \mathrm{E}[\mathrm{tr}(I_n - P_F)(\boldsymbol{y}-\boldsymbol{\eta}_0)(\boldsymbol{y}-\boldsymbol{\eta}_0)'] \\
&= \mathrm{tr}(I_n - P_F)\sigma_0^2 = (n-k-1)\sigma_0^2. \qquad (2.5)
\end{aligned}
$$

The theorem follows from (2.1), (2.3) and (2.5).

It is well known (see, e.g. Allen (1971, 1974), Hocking (1972), Haga et al. (1973)) that C_{cv} can be written as

$$C_{cv} = \sum_{\alpha=1}^{n}(y_\alpha - \hat{y}_{(-\alpha)J})^2 = \sum_{\alpha=1}^{n}\left(\frac{y_\alpha - \hat{y}_{\alpha J}}{1-c_\alpha}\right)^2,$$

where c_α is the (α, α)th element of P_J. Therefore, we have

$$\begin{aligned} C_{cv} &= \sum_{\alpha=1}^{n} (y_\alpha - \hat{y}_{\alpha J})^2 \left\{ 1 + \frac{c_\alpha}{1 - c_\alpha} \right\}^2 \\ &= \sum_{\alpha=1}^{n} (y_\alpha - \hat{y}_{\alpha J})^2 + (\boldsymbol{y} - \hat{\boldsymbol{y}}_J)' D_a (\boldsymbol{y} - \hat{\boldsymbol{y}}_J) \\ &= s_J^2 + \hat{B}_{cv}, \end{aligned} \tag{2.6}$$

where

$$\begin{aligned} & D_a = \mathrm{diag}(a_1, \ldots, a_n), \\ & a_\alpha = 2\frac{c_\alpha}{1 - c_\alpha} + \left(\frac{c_\alpha}{1 - c_\alpha} \right)^2, \quad \alpha = 1, \ldots, n, \\ & \hat{B}_{cv} = (\boldsymbol{y} - \hat{\boldsymbol{y}}_J)' D_a (\boldsymbol{y} - \hat{\boldsymbol{y}}_J). \end{aligned}$$

Theorem 2.2 *The biase B_{cv} when we estimate R_{pe} by the cross-validation criterion C_{cv} can be expressed as*

$$\begin{aligned} B_{cv} &= \mathrm{E}_0(C_{cv}) - R_{pe} \\ &= \left(\sum_{\alpha=1}^{n} \frac{c_\alpha^2}{1 - c_\alpha} \right) \sigma_0^2 + \tilde{\delta}_J^2, \end{aligned} \tag{2.7}$$

where $\tilde{\delta}_J^2 = \{(I_n - P_J)\boldsymbol{\eta}_0\}' D_a \{(I_n - P_J)\boldsymbol{\eta}_0\}$. In particular, when the true model is contained in the model M_J, we have

$$B_{cv} = \left(\sum_{\alpha=1}^{n} \frac{c_\alpha^2}{1 - c_\alpha} \right) \sigma_0^2. \tag{2.8}$$

Proof

We can write $\hat{B}_{cv}$ as follows.

$$\begin{aligned} \hat{B}_{cv} &= \{(I_n - P_J)\boldsymbol{y}\}' D_a \{(I_n - P_J)\boldsymbol{y}\} \\ &= \mathrm{tr}\{(I_n - P_J)\boldsymbol{y}\}' D_a \{(I_n - P_J)\boldsymbol{y}\} \\ &= \mathrm{tr} D_a \{(I_n - P_J)\boldsymbol{y}\}\{(I_n - P_J)\boldsymbol{y}\}' \\ &= \mathrm{tr} D_a \{(I_n - P_J)\{(\boldsymbol{y} - \boldsymbol{\eta}_0) + \boldsymbol{\eta}_0\} \\ &\qquad \times \{(\boldsymbol{y} - \boldsymbol{\eta}_0) + \boldsymbol{\eta}_0\}'(I_n - P_J)\}. \end{aligned}$$

Therefore we have

$$\begin{aligned}
\mathrm{E}(\hat{B}_{cv}) &= \mathrm{tr}D_a\{(I_n - P_J)\{\sigma_0^2 I_n + \boldsymbol{\eta}_0\boldsymbol{\eta}_0'\}(I_n - P_J)\\
&= \sum_{\alpha=1}^{n}\left\{2\frac{c_\alpha}{1-c_\alpha} + \left(\frac{c_\alpha}{1-c_\alpha}\right)^2\right\}(1-c_\alpha)\sigma_0^2 + \tilde{\delta}_J^2\\
&= \left\{2(j+1) + \sum_{\alpha=1}^{n}\frac{c_\alpha^2}{1-c_\alpha}\right\}\sigma_0^2 + \tilde{\delta}_J^2.
\end{aligned}$$

The required result is obtained from the above result, Lemma 2.1 and (2.6).

It is natural to assume that $c_i = O(n^{-1})$, since $0 \le c_i$ and $\sum_{i=1}^n c_i = k$. Then

$$\sum_{j=1}^{n}\frac{c_j^2}{1-c_j} \le \frac{1}{1-\bar{c}}\sum_{i=1}^{n} c_i^2 = O(n^{-1}).$$

This implies that $B_{cv} = O(n^{-1})$ and hence C_{cv} is asymptotically unbiased when the true model is contained in the candidate model. On the other hand, C_{pe} is exactly unbiased under a weaker condition, i.e. when the true model is contained in the full model.

3 Relation of C_{pe} with C_p and C_{mp}

We can write C_p criterion (Mallows (1973, 1995)) as

$$\begin{aligned}
C_p &= \frac{s_J^2}{\hat{\sigma}^2} + 2(j+1)\\
&= (n-k-1)\frac{s_J^2}{s_F^2} + 2(j+1), \qquad (3.1)
\end{aligned}$$

where $\hat{\sigma}^2$ is the usual unbiased estimator of σ^2 under the full model, and is given by $\hat{\sigma}^2 = s_F^2/(n-k-1)$. The criterion was proposed as an estimator for the standardized mean square errors in prediction given by

$$\tilde{R}_{pe} = \sum_{\alpha=1}^{n}\mathrm{E}_0[\frac{1}{\sigma_0^2}(z_\alpha - \hat{y}_{\alpha J})^2] = \sum_{\alpha=1}^{n}\mathrm{E}_0[\frac{1}{\sigma_0^2}(\eta_{\alpha 0} - \hat{y}_\alpha)^2] + n. \qquad (3.2)$$

Mallows (1973) originally proposed

$$\frac{s_J^2}{\hat{\sigma}^2} + 2(j+1) - n$$

as an estimator for the first term in the last expression of (3.2). In this paper we call (3.1) C_p criterion. Fujikoshi and Satoh (1997) proposed a modified C_p criterion defined by

$$C_{mp} = (n-k-3)\frac{s_J^2}{s_F^2} + 2(j+2). \tag{3.3}$$

They show that C_{mp} is an exact unbiased estimator for $\tilde{R}_{pe}$ when the true model is contained in the full model and the errors are normally distributed. As we have seen, C_{pe} has the same property for its target measure R_{pe}. However, it may be noted that the normality assumption is not required for C_{pe} criterion. Among these three criteria, there are the following close relationships given by

$$C_{pe} = \frac{s_F^2}{n-k-1}C_p, \tag{3.4}$$

$$C_{mp} = C_p + 2\left(1 - \frac{s_J^2}{s_F^2}\right). \tag{3.5}$$

This means that these three criteria choose the same model while they have different properties such that they are unbiased estimators for the target measures R_{pe} and $\tilde{R}_{pe}$, respectively.

4 Modifications of multiple correlation coefficient

Let R be the multiple correlation coefficient between y and $(x_1, \ldots, x_j)$ which may be defined by

$$R^2 = 1 - s_J^2/s_y^2.$$

As an alternative criterion for selecion variables, we sometime encounter the multiple correlation coefficient $\tilde{R}$ adjusted for the degree of freedom given by

$$\begin{aligned}\tilde{R}^2 &= 1 - \frac{s_J^2/(n-j-1)}{s_y^2/(n-1)} \\ &= 1 - \frac{n-1}{n-j-1}(1-R^2) \end{aligned} \tag{4.1}$$

The criterion chooses the model which $\tilde{R}^2$ is maximized. We consider a transformed criterion defined by

$$\tilde{C}_{dc} = (1-\bar{R}^2)s_y^2 = \frac{n-1}{n-j-1}s_J^2, \tag{4.2}$$

which may be regarded as an estimator of R_{pe}. However, as we shall see lator, $\tilde{C}_{dc}$ is not unbiased even when the true model is contained in the model M_J. In this paper we propose an adjusted multiple correlation coefficient given by

$$\bar{R}^2 \;=\; 1 - \frac{n+j+1}{n-j-1}(1-R^2) \tag{4.3}$$

whose determination coefficient is defined by (1.8). The unbiasedness property is given in the following theorem.

Theorem 4.1 *Consider an adjusted multiple correlation coefficient $\tilde{R}_a$ defined by*

$$\tilde{R}_a^2 = 1 - a(1-R^2)$$

and the corresponding determination coefficient defined by

$$\begin{aligned}\tilde{C}_{dc;a} &= s_y^2(1-\tilde{R}_a^2)\\ &= as_J^2\end{aligned}$$

as in (4.2)*, where a is a constant depending the sample size n. Then we have*

$$\mathrm{E}(\tilde{C}_{dc;a}) = R_{pe} + B_{dc;a},$$

where

$$B_{dc;a} = \{(a-1)(n-j-1) - 2(j+1)\}\sigma_0^2 + (a-1)\delta_J^2$$

with $\delta_J^2 = \boldsymbol{\eta}_0'(I_n - P_J)\boldsymbol{\eta}_0$. Further, if the true model is contained in the model M_J, $\delta_J^2 = 0$, and $\tilde{C}_{dc;a}$ is an unbiased estimator if and only if

$$a = \frac{n+j+1}{n-j-1}.$$

Proof

We decompose $\tilde{C}_{dc;a}$ as

$$\tilde{C}_{dc;a} = s_J^2 + (a-1)s_J^2.$$

Applying (2.1) and (2.3) in Lemma 2.1 to each term of the decomposition,

$$\mathrm{E}_0(\tilde{C}_{dc;a}) = R_{pe} - 2(j+1)\sigma_0^2 + (a-1)\{(n-j-1)\sigma_0^2 + \delta_J^2\}$$

which implies the first result and hence the remeinder result.

In general, we have

$$\mathrm{E}_0(\bar{C}_{dc}) = R_{pe} + \frac{2(j+1)}{n-j-1}\delta_J^2, \tag{4.4}$$

and the order of the bias is $O(n^{-1})$. Haga et al. (1973) proposed an alternative adjusted multiple correlation coefficient $\hat{R}$ defined by

$$\begin{aligned} \hat{R}^2 &= 1 - \frac{(n+j+1)s_J^2/(n-j-1)}{(n+1)s_y^2/(n-1)} \\ &= 1 - \frac{(n-1)(n+j+1)}{(n+1)(n-j-1)}(1-R^2). \end{aligned}$$

The corresponding determination coefficient is

$$\hat{C}_{dc} = \frac{(n-1)(n+j+1)}{(n+1)(n-j-1)}s_J^2.$$

From (2.4) we can see (see Haga et al.(1973)) that if the true model is cotained in the model M_J, then

$$\mathrm{E}[(n+j+1)s_J^2/(n-j-1)] = R_{pe} = R_{pe}(J).$$

In particular, if J is the empty set ϕ,

$$\mathrm{E}[(n+1)s_y^2/(n-1)] = R_{pe}(\phi).$$

Theorem 3.1 implies that $\hat{C}_{dc}$ is not unbiased as an estimator of R_{pe} even when the true model is contained in the model M_J. In fact

$$\begin{aligned} \mathrm{E}_0(\hat{C}_{dc}) &= R_{pe} - \frac{n+2}{n+1}\sigma_0^2 + \frac{2jn}{(n+1)(n-j-1)}\delta_J^2 \\ &= R_{pe} - \frac{n+2}{n+1}\sigma_0^2, \ \text{if} M_0 \text{ is contained in } M_J. \end{aligned}$$

5 Multivariate version of C_{pe}

In this section we consider a multivariate linear regression model of p response variables $y_1, \ldots, y_p$ and k explanatory variables $x_1, \ldots, x_k$. Suppose that we have an sample of $\boldsymbol{y} = (y_1, \ldots, y_p)'$ and $\boldsymbol{x} = (x_1, \ldots, x_k)'$ of size n given by

$$\boldsymbol{y}_\alpha = (y_{\alpha 1}, \ldots, y_{\alpha p})', \quad \boldsymbol{x}_\alpha = (x_{\alpha 1}, \ldots, x_{\alpha k})'; \ \alpha = 1, \ldots, n.$$

A multivariate linear model is given by

$$\begin{aligned} M_F: \ Y &= (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)' \\ &= (\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n)'(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta})' + (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)' \\ &= X\mathcal{B} + \mathcal{E}, \end{aligned} \tag{5.1}$$

where the error terms $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are mutually independent, and each of them has the same mean vector $\mathbf{0}$ and the same unknown covariance matrix Σ. The linear regression model based on the subset of the first j explanatory variables can be expressed as

$$M_J; \ Y = X_J\mathcal{B}_J + \mathcal{E}, \tag{5.2}$$

where $\mathcal{B}_J = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_j)'$. The true model for Y is assumed to be

$$\begin{aligned} M_0; \ Y &= (\boldsymbol{\eta}_{10}, \ldots, \boldsymbol{\eta}_{n0})' + (\boldsymbol{\varepsilon}_{10}, \ldots, \boldsymbol{\varepsilon}_{n0})' \\ &= \mathcal{Y}_0 + \mathcal{E}_0, \end{aligned} \tag{5.3}$$

where the error terms $\boldsymbol{\varepsilon}_{10}, \ldots, \boldsymbol{\varepsilon}_{n0}$ are mutually independent, and each of them has the same mean vector $\mathbf{0}$ and the same covariance matrix Σ_0.

Let $\hat{\boldsymbol{y}}_{\alpha J}$ be the best linear unbiased estimator of $\boldsymbol{\eta}_{\alpha 0}$ under a candidate model M_J. The criterion (1.4) for goodness of a fitted candidate model is extended as

$$\begin{aligned} R_{pe} &= \sum_{\alpha=1}^{n} \mathrm{E}_0[(\boldsymbol{z}_\alpha - \hat{\boldsymbol{y}}_{\alpha J})'(\boldsymbol{z}_\alpha - \hat{\boldsymbol{y}}_{\alpha J})] \\ &= \mathrm{E}_0[\mathrm{tr}(Z - \hat{Y}_J)'(Z - \hat{Y}_J)], \end{aligned} \tag{5.4}$$

where $\hat{Y}_J = (\hat{\boldsymbol{y}}_{1J}, \ldots, \hat{\boldsymbol{y}}_{nJ})'$, and $Z = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)'$ is independent of the observation matrix is distributed as in (5.3), and E_0 denotes the expectation with respect to the true model (5.3). Then we can express R_{pe} as

$$\begin{aligned} R_{pe} &= \sum_{\alpha=1}^{n} \mathrm{E}_0[(\boldsymbol{\eta}_{\alpha 0} - \hat{\boldsymbol{y}}_{\alpha J})'(\boldsymbol{\eta}_{\alpha 0} - \hat{y}_{\alpha J})] + n\mathrm{tr}\Sigma_0 \\ &= \mathrm{E}[\mathrm{tr}(\mathcal{Y}_0 - \hat{Y})'(\mathcal{Y}_0 - \hat{Y})] + n\mathrm{tr}\Sigma_0. \end{aligned} \tag{5.5}$$

In a cross-validiation for the multivariate prediction error (5.5), $\boldsymbol{y}_\alpha$ is predicted by the predictor $\hat{\boldsymbol{y}}_{(-\alpha)J}$ based on the data set obtained by removing the αth observation $(\boldsymbol{y}_\alpha, \boldsymbol{x}_\alpha)$, and R_{pe} is estemated by

$$C_{cv} = \sum_{\alpha=1}^{n} (\boldsymbol{y}_\alpha - \hat{\boldsymbol{y}}_{(-\alpha)J})'(\boldsymbol{y}_\alpha - \hat{\boldsymbol{y}}_{(-\alpha)J}). \tag{5.6}$$

By the same way as in the univariate case, we have

$$
\begin{aligned}
C_{cv} &= \sum_{\alpha=1}^{n} (\boldsymbol{y}_\alpha - \hat{\boldsymbol{y}}_{(-\alpha)J})'(\boldsymbol{y}_\alpha - \hat{\boldsymbol{y}}_{(-\alpha)J}) \\
&= \sum_{\alpha=1}^{n} \left(\frac{1}{1-c_\alpha} \right)^2 (\boldsymbol{y}_\alpha - \hat{\boldsymbol{y}}_{\alpha J})'(\boldsymbol{y}_\alpha - \hat{\boldsymbol{y}}_{\alpha J}).
\end{aligned}
$$

Now, our main interest is an extension of C_{pe} to multivariate case. Let S_J and S_F be the matrices of sums of squares and products under the candidate model M_J and the full model M_F, respectively. These matrices are given by

$$
\begin{aligned}
S_J &= (Y - \hat{Y}_J)'(Y - \hat{Y}_J) = Y'(I_n - P_J)Y, \\
S_F &= (Y - \hat{Y}_F)'(Y - \hat{Y}_F) = Y'(I_n - P_F)Y,
\end{aligned}
$$

where

$$
\hat{Y}_J = X_J(X_J'X_J)^{-1}Y = P_JY, \quad \hat{Y}_F = X_F(X_F'X_F)^{-1}Y = P_FY.
$$

As an estinator of (5.5), we consider

$$
C_{pse} = \mathrm{tr}S_J + \frac{2(j+1)}{n-k-1}\mathrm{tr}S_F. \tag{5.7}
$$

Then the following result is demonstrated.

Theorem 5.1 *Suppose that the true model M_0 is contained in the full model M_F. the C_{pe} in* (5.7) *is an unbiased estimator of the multivariate prediction error R_{pe} in* (5.5).

Proof

By an argument similar to one as in Lemma 2.1, we can show that

$$
\begin{aligned}
\mathrm{E}_0[(Z - \hat{Y}_J)'(Z - \hat{Y}_J)] &= (n + j + 1)\Sigma_0 + \Delta_J, \\
\mathrm{E}_0[(Y - \hat{Y}_J)'(Y - \hat{Y}_J)] &= (n - j - 1)\Sigma_0 + \Delta_J,
\end{aligned}
$$

where $\Delta_J = \mathcal{Y}_0'(I_n - P_J)\mathcal{Y}_0$. Further, since the true model is contained in the full model,

$$
\mathrm{E}(S_F) = (n - k - 1)\Sigma_0,
$$

which implies the required result.

The C_p and C_{mp} criteria in univariate case have been extended（Fujikoshi and Satoh (1997)）as

$$C_p = (n-k-1)\mathrm{tr}S_J S_F^{-1} + 2p(j+1),$$
$$C_{mp} = (n-k-p-2)\mathrm{tr}S_J S_F^{-1} + 2p(j+1) + p(p+1),$$

respectively. The results in Section 2 may be extended similarly, but its details are omitted here.

6 Numerical example

Consider Hald's example on examining the heat generated during the hardening of Portland cement. The following variables were measured (see, e.g., Flury and Riedwy (1988)).

x_1 = amount of tricalcium aluminate,
x_2 = amount of tricalcium silicate,
x_3 = amount of tetracalcuim alumino ferrite
x_4 = amount of dicalcium silicate,
y = heat evolved in calories.

The observations with the sample size $n = 13$ are given in Table 6.1.

Table 6.1. Data of the cement hardning example

α	$x_{\alpha 1}$	$x_{\alpha 2}$	$x_{\alpha 3}$	$x_{\alpha 4}$	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	29	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

Now we consider all the candidate models except the constant model, and denote the models obtained by using $\{x_1\}, \{x_1, x_2\}, \ldots$, by $M_1, M_{1,2}, \ldots$, respectively. The number of such models is

$$ {}_4C_1 + {}_4C_2 + {}_4C_3 + {}_4C_4 - 1 = (1+1)^4 - 1 = 2^4 - 1 = 15. $$

For each of all the candidate models, we computed the values of the following basic quantities and criteria in Table 6.2:

R^2; squares of multiple correlation coefficients,
$\hat{\sigma}^2$; the usual unbiased estimator of σ^2,
C_p; Mallows C_p criterion,
C_{mp}; modified C_p criterion,
C_{cv}; cross validation criterion,
C_{pe}; prediction error criterion,
C_{dc}; determination coefficients,
$\hat{C}_{dc}$; modified determination coefficient,
$\bar{C}_{dc}$; adjusted determination coefficient.

Table 6.2. The values of R^2, $\hat{\sigma}^2$, C_p, C_{mp}, C_{cv}, C_{pe}, C_{dc}, $\hat{C}_{dc}$ and $\bar{C}_{dc}$

models	R^2	$\hat{\sigma}^2$	C_p	C_{mp}	C_{cv}	C_{pe}	C_{cd}	$\hat{C}_{dc}$	$\bar{C}_{dc}$
M_1	0.5339	115.1	215.5	164.7	1699.6	1289.6	0.5084	0.5447	0.6355
M_2	0.6663	82.4	155.5	119.6	1202.1	930.3	0.3641	0.3901	0.4551
M_3	0.2859	176.31	328.2	249.1	2616.4	1963.3	0.7791	0.8347	0.9738
M_4	0.6745	80.4	151.7	116.8	1194.2	907.8	0.3551	0.3804	0.4438
M_{12}	0.9787	5.8	15.7	15.3	93.9	93.8	0.0256	0.0292	0.0341
M_{13}	0.5482	122.7	211.1	61.8	2218.1	1263.0	0.5422	0.6197	0.7229
M_{14}	0.9725	7.5	18.5	17.4	121.2	110.7	0.0330	0.0378	0.0441
M_{23}	0.8470	41.5	75.4	60.1	701.7	451.3	0.1836	0.2098	0.2448
M_{24}	0.6801	86.9	151.2	116.9	1461.8	904.8	0.3839	0.4388	0.5119
M_{34}	0.9353	17.6	35.4	30.0	294.0	211.6	0.0777	0.0888	0.1035
M_{123}	0.9823	5.3	16.0	16.0	90.0	96.0	0.0236	0.0287	0.0335
M_{124}	0.9823	5.3	16.0	16.0	85.4	95.8	0.0236	0.0286	0.0334
M_{134}	0.9813	5.6	16.5	16.4	94.5	98.7	0.0250	0.0303	0.0354
M_{234}	0.9728	8.2	20.3	19.3	146.9	121.7	0.0362	0.0440	0.0513
M_{1234}	0.9824	6.0	18.0	18.0	110.3	107.7	0.0264	0.0340	0.0397

All the three criteria C_p, C_{mp} and C_{pe} choose the model M_{12} as an optimum model. However, the other four criteria C_{cv}, C_{dc}, $\hat{C}_{dc}$ and $\bar{C}_{dc}$ choose a larger model M_{124} which contains M_{12} as an optimum model. Each of the three criteria C_{dc}, $\hat{C}_{dc}$ and $\bar{C}_{dc}$ are almost the same for models M_{123} and M_{124}. As being noted in Section 3 the three criteria C_p, C_{mp} and C_{pe} always choose the same model as an optimum model. In general, the criteria C_{cv}, C_{dc}, $\hat{C}_{dc}$ and $\bar{C}_{dc}$ shall have a tendancy of choosing a large model in the comparison with the criteria C_p, C_{mp} and C_{pe}.

References

[1] Akaike, H. (1973). Informaiton theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, (B. N. Petrov and F.Csáki,eds.), 267-81, Budapest: Akadémia Kiado.

[2] Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, **13**, 469-475.

[3] Allen, D. M. (1974). The relationship between variable selection and data augumentation, and a method for prediction. *Technometrics*, **16**,

[4] Flury, B. and Riedwy, H. (1988). *Multivariate Statistics - A Practical Approach* -. Chapman and Hall.

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

[6] Haga, Y., Takeuchi, K. and Okuno, C. (1973). New criteria for selecting of variables in regression model. *Quality (Hinshitsu, J. of the Jap. Soc. for Quality Control)*, **6**, 73-78 (in Japanese).

[7] Hocking, R. R. (1972). Criteria for selecting of a subset regression; which one should be used. *Technometrics*, **14**, 967-970.

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

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

[10] Stone, M. (1974). Cross-validatory choice and assesment of statistical predictions (with Discussion). *J. R. Statist. Soc.*, **B**, **36**, 111-147.