Ann Inst Stat Math (2011) 63:387–403
DOI 10.1007/s10463-009-0233-5

Prediction error criterion for selecting variables in a linear regression model

Yasunori Fujikoshi · Tamio Kan · Shin Takahashi · Tetsuro Sakurai

Received: 12 September 2007 / Revised: 1 September 2008 / Published online: 30 April 2009

Abstract Several criteria, such as CV, C_p, AIC, CAIC, and MAIC, are used for selecting variables in linear regression models. It might be noted that C_p has been proposed as an estimator of the expected standardized prediction error, although the target risk function of CV might be regarded as the expected prediction error R_{PE}. On the other hand, the target risk function of AIC, CAIC, and MAIC is the expected log-predictive likelihood. In this paper, we propose a prediction error criterion, PE, which is an estimator of the expected prediction error R_{PE}. Consequently, it is also a competitor of CV. Results of this study show that PE is an unbiased estimator when the true model is contained in the full model. The property is shown without the assumption of normality. In fact, PE is demonstrated as more faithful for its risk function than CV. The prediction error criterion PE is extended to the multivariate case. Furthermore, using simulations, we examine some peculiarities of all these criteria.

Keywords Prediction error criterion · Linear regression models · Selection of variables · Risk function · Selection criteria

1 Introduction

With a linear regression model, we seek to predict or describe a response variable y using the full set of explanatory variables $x_1, \ldots, x_k$ or its subsets. Assume n

Y. Fujikoshi (✉) · T. Sakurai
Department of Mathematics, Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
e-mail: fujikoshi_y@yahoo.co.jp

T. Kan · S. Takahashi
Esumi Co. Ltd., Nakano F bldg. 8F, 4-44-18 Honcho, Nakano-ku, Tokyo 164-0012, Japan

observations on y and $\boldsymbol{x} = (x_1, \ldots, x_k)'$ denoted by $y_\alpha, \boldsymbol{x}_\alpha = (x_{\alpha 1}, \ldots, x_{\alpha k})'; \alpha = 1, \ldots, n$. As a model based on a subvector of $\boldsymbol{x}$, without loss of generality, we might consider the model based on the first $j (\leq k)$ explanatory variables $x_1, \ldots, x_j$, which is expressed as

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

where the coefficients $\beta_0, \ldots, \beta_j$ are unknown, and the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are mutually independent and have the same mean 0 and the same unknown variance σ^2. This model M_J is also called as a candidate model. The linear regression model, including all the explanatory variables, is given as

$$M_F : \ y_\alpha = \beta_0 + \beta_1 x_{\alpha 1} + \cdots + \beta_k x_{\alpha k} + \varepsilon_\alpha, \quad \alpha = 1, \ldots, n, \tag{2}$$

where the coefficients $\beta_0, \ldots, \beta_k$ are unknown parameters, and the error terms $\varepsilon_1, \ldots, \varepsilon_n$ have the same distributions as in (1). The model (2) is called the full model. We assume that the true model for $y_\alpha, \alpha = 1, \ldots, n$ is as follows:

$$M_* : \ y_\alpha = \eta_\alpha + \varepsilon_\alpha, \quad \alpha = 1, \ldots, n, \tag{3}$$

where the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are mutually independent; each of them has the same mean 0 and the same variance σ_0^2.

For the selection of the best model from a collection of candidate models specified by linear regression of y on subvectors of $\boldsymbol{x}$, it is important to define a measure of goodness of a candidate model, i.e., or to define what the target model is. In Sect. 2, we provide a brief review of such measures and their estimators.

In this paper, we consider the expected prediction error, given as

$$R_{\mathrm{PE}} = \sum_{\alpha=1}^{n} \mathrm{E}_{\boldsymbol{y}}^* \mathrm{E}_{\boldsymbol{z}}^* [(z_\alpha - \hat{y}_{\alpha J})^2], \tag{4}$$

as a measure of goodness of fit for a candidate model M_J, where $\hat{y}_{\alpha J}$ is the usual unbiased estimator of η_α under M_J. Here $\boldsymbol{z} = (z_1, \ldots, z_n)'$ is a future observation vector; it has the same distribution as $\boldsymbol{y} = (y_1, \ldots, y_n)'$ in (3) and is independent of $\boldsymbol{y}$. Furthermore, $\mathrm{E}_{\boldsymbol{y}}^*$ and $\mathrm{E}_{\boldsymbol{z}}^*$ respectively denote the expectations with respect to $\boldsymbol{y}$ and $\boldsymbol{z}$ when they are distributed according to the true model M_*. The measure R_{PE} is also regarded as a risk function for M_J. It is readily apparent that

$$R_{\mathrm{PE}} = \sum_{\alpha=1}^{n} \mathrm{E}_{\boldsymbol{y}}^* [(\eta_\alpha - \hat{y}_{\alpha J})^2] + n\sigma_0^2. \tag{5}$$

Therefore, the target risk function is fundamentally identical to the first term of the right-hand side in (5). Mallows (1973) considered the expected standardized prediction error $R_{\mathrm{PE}}/\sigma_0^2$ as the target risk function, as described in Sect. 2.

The cross-validation method (see, e.g. Stone 1974) is closely related to a criterion based on the expected prediction error R_{PE}. In fact, the method predicts y_α using the usual unbiased estimator $\hat{y}_{(-\alpha)J}$ based on the data set obtained by removing the αth observation $(y_\alpha, \boldsymbol{x}'_\alpha)$. The CV criterion is defined as

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

In fact, CV can be regarded as an estimator of R_{PE}. The selection method is to select the model for which CV is minimized.

As a competitor of CV, we propose a prediction error criterion

$$\mathrm{PE} = s_J^2 + \frac{2(j+1)}{n-k-1} s_F^2,$$

where s_J^2 and s_F^2 respectively represent the sums of squares of residuals in a candidate model M_J and the full model M_F. In Sect. 2, we present a brief review of C_p, AIC and their modifications. We shall give a relationship of PE with C_p (Mallows 1973) and its modification MC_p (Fujikoshi and Satoh 1997).

In Sect. 3, we examine the unbiased properties of CV and PE as an estimator for their target risk function R_{PE}. Results of the investigation reveal that CV is asymptotically unbiased, whereas PE is exactly unbiased when the true model is contained in the full model. Both C_p and MC_p are closely related to PE because the target risk function for C_p and MC_p might be considered as the expected standardized prediction error $R_{\mathrm{PE}}/\sigma_0^2$. We note that PE and C_p select the same model as a best model.

In Sect. 4, we describe a multivariate extension of PE. In Sect. 5, the unbiased properties of PE and CV as estimators of the target risk function are also examined through simulation experiments. Note that {CV, PE}, $\{C_p, MC_p\}$ and {AIC, CAIC, MAIC} have their own target risk functions. However, all these criteria may be used as a criterion for selection of the true model, more precisely the minimal model which includes the true model or its approximation. We give the relative performance of selection of the true model by simulation experiments.

2 Brief review of C_p, AIC, and their modifications

We can write the C_p criterion (Mallows 1973, 1975) for M_J as

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

where $\hat{\sigma}^2$ is the usual unbiased estimator of σ^2 under the full model M_F; it is given as $\hat{\sigma}^2 = s_F^2/(n-k-1)$. The criterion can be considered as an estimator for the expected standardized expected prediction error given as

$$\tilde{R}_{\mathrm{PE}} = \sum_{\alpha=1}^{n} \mathrm{E}_{\mathbf{y}}^{*}\mathrm{E}_{\mathbf{z}}^{*}\left[\frac{1}{\sigma_0^2}(z_\alpha - \hat{y}_{\alpha J})^2\right]$$
$$= \sum_{\alpha=1}^{n} \mathrm{E}_{\mathbf{y}}^{*}\left[\frac{1}{\sigma_0^2}(\eta_\alpha - \hat{y}_\alpha)^2\right] + n. \tag{6}$$

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 (6). Fujikoshi and Satoh (1997) proposed a modified C_p criterion defined as

$$MC_p = (n-k-3)\frac{s_J^2}{s_F^2} + 2(j+2).$$

They showed that MC_p is an exact unbiased estimator for $\tilde{R}_{\mathrm{PE}}$ when the true model is contained in the full model and the errors are normally distributed. As described in Sect. 3, PE has the same property for its target risk R_{PE}. Furthermore, it might be noted that the normality assumption is not required for the PE criterion. Among these three criteria, the following close relationships pertain:

$$\mathrm{PE} = \frac{s_F^2}{n-k-1}C_p,$$
$$MC_p = C_p - 2\left(\frac{s_J^2}{s_F^2} - 1\right) \leq C_p. \tag{7}$$

The "best" model is defined as the model which minimizes the target risk function considered. For this reason, it depends on the criterion used. The first result (7) shows that PE and C_p select the same model as a best model, although their target risk functions are different. On the other hand, a model selection criterion is also used to find models which almost all have the same values as the best model. Related to the later use, it is important that a criterion is faithful for its risk function. Especially, for example, a criterion is required to be exactly or approximately unbiased for its risk function.

In Sect. 3, we show that PE is an unbiased estimator of R_{PE} when the true model is contained in the full model. Therefore, the degree of differences of PE corresponds to one difference of R_{PE}. Such an unbiased property does not hold for C_p. In fact, from (7), C_p overestimates $\tilde{R}_{\mathrm{PE}}$ when the true model is contained in the full model. The errors are normally distributed because MC_p is an unbiased estimator of $\tilde{R}_{\mathrm{PE}}$.

We can use AIC when the errors in M_J are normally distributed (Akaike 1973). Akaike (1973) used the expected predictive likelihood defined as

$$R_A = \mathrm{E}^*_{\boldsymbol{y}}\mathrm{E}^*_{\boldsymbol{z}}\left[-2\log\left\{\prod_{\alpha=1}^{n}\left(2\pi\hat{\sigma}_J^2\right)^{-1/2}\exp\left(-\frac{1}{2\hat{\sigma}_J^2}(z_\alpha-\hat{y}_{\alpha J})^2\right)\right\}\right]$$
$$= \mathrm{E}^*_{\boldsymbol{y}}\left[n\log\hat{\sigma}_J^2 + n(\log 2\pi+1)\right] + B_A \tag{8}$$

as a measure of goodness of a fitted model M_J, where

$$B_A = \mathrm{E}^*_{\boldsymbol{y}}\mathrm{E}^*_{\boldsymbol{z}}\left[\sum_{\alpha=1}^{n}\frac{1}{\hat{\sigma}_J^2}\left(z_\alpha-\hat{y}_{\alpha,J}\right)^2\right]-n,$$

and $\hat{\sigma}_J^2 = s_j^2/n$ is the maximum likelihood estimate of σ^2 under M_J. Then AIC is defined as

$$\mathrm{AIC} = n\log\hat{\sigma}_J^2 + n(\log 2\pi+1) + 2(j+2),$$

which is an asymptotic unbiased estimator of R_A when the true model is contained in the candidate model M_J. Sugiura (1978) and Bedrick and Tsai (1994) proposed

$$\mathrm{CAIC} = n\log\hat{\sigma}_J^2 + n(\log 2\pi+1) + \frac{2n(j+2)}{n-j-3}$$
$$= \mathrm{AIC} + \frac{2(j+2)(j+3)}{n-j-3},$$

which is an unbiased estimator of R_A when the true model is contained in the candidate model M_J. Relaxing the restriction that the true model is included in the candidate model M_J, Fujikoshi and Satoh (1997) proposed a modification

$$\mathrm{MAIC} = \mathrm{CAIC} + 2(Q-1)(j+2-Q),$$

where $Q = \{s_F^2/(n-k-1)\}/\{s_J^2/(n-j-1)\}$. In fact, MAIC is known to have better estimator than AIC or CAIC on unbiasedness when the true model is not always included in a candidate model M_J, but is contained in the full model M_F.

Davies et al. (2006) showed that MC_p and CAIC achieve minimum variance within the class of unbiased estimators.

3 Unbiasedness of CV and PE

Writing the model M_J in (1) as in matrix form, we have

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

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

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

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

We can write R_{PE} as

$$\begin{aligned}R_{\mathrm{PE}} &= \mathrm{E}^*_{\boldsymbol{y}}\mathrm{E}^*_{\boldsymbol{z}}[(\boldsymbol{z} - \hat{\boldsymbol{y}}_J)'(\boldsymbol{z} - \hat{\boldsymbol{y}}_J)] \\ &= \mathrm{E}^*_{\boldsymbol{y}}[(\boldsymbol{y} - \hat{\boldsymbol{y}}_J)'(\boldsymbol{y} - \hat{\boldsymbol{y}}_J)] + B_{\mathrm{PE}} \\ &= \mathrm{E}^*_{\boldsymbol{y}}[s_J^2] + B_{\mathrm{PE}}, \end{aligned} \tag{9}$$

where

$$B_{\mathrm{PE}} = \mathrm{E}^*_{\boldsymbol{y}}\mathrm{E}^*_{\boldsymbol{z}}[(\boldsymbol{z} - \hat{\boldsymbol{y}}_J)'(\boldsymbol{z} - \hat{\boldsymbol{y}}_J) - (\boldsymbol{y} - \hat{\boldsymbol{y}}_J)'(\boldsymbol{y} - \hat{\boldsymbol{y}}_J)]. \tag{10}$$

Using (9) it is readily apparent that B_{PE} is the bias term when we estimate R_{PE} using s_J^2.

Theorem 1 *The bias term B_{PE} in (10) can be evaluated as*

$$B_{\mathrm{PE}} = 2(j+1)\sigma_0^2.$$

Furthermore, if the true model M_* is contained in the full model M_F, the criterion PE is an unbiased estimator for R_{PE}.

Proof We have

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

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'$ and $\delta_J^2 = \boldsymbol{\eta}'(I_n - P_J)\boldsymbol{\eta}$. The result is obtained using $P_J^2 = P_J$, and

$$\begin{aligned}\mathrm{E}^*_{\boldsymbol{y}}[(\boldsymbol{y} - \boldsymbol{\eta})' P_J (\boldsymbol{y} - \boldsymbol{\eta})] &= \mathrm{E}^*_{\boldsymbol{y}}[\mathrm{tr} P_J (\boldsymbol{y} - \boldsymbol{\eta})(\boldsymbol{y} - \boldsymbol{\eta})'] \\ &= \mathrm{tr} P_J \sigma_0^2 = (j+1)\sigma_0^2.\end{aligned}$$

Similarly, we have

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

These imply the first result. The second result follows the fact that $S_F^2/(n-k-1)$ is an unbiased estimator of σ_0^2. Although this result is well known, we describe a derivation. It is worth noting that $s_F^2=\boldsymbol{y}'(I_n-P_F)\boldsymbol{y}$, where $P_F=X(X'X)^{-1}X'$. Because the true model M_* is contained in the full model M_F, $P_F\boldsymbol{\eta}=\boldsymbol{\eta}$. We have

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

□

Actually, CV is well known (e.g., Allen 1971, 1974; Hocking 1972; Haga et al. 1973) to be expressible as

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

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

$$\begin{aligned}
\mathrm{CV} &= \sum_{\alpha=1}^{n}(y_\alpha-\hat{y}_{\alpha J})^2\left\{1+\frac{c_\alpha}{1-c_\alpha}\right\}^2 \\
&= \sum_{\alpha=1}^{n}(y_\alpha-\hat{y}_{\alpha J})^2+(\boldsymbol{y}-\hat{\boldsymbol{y}}_J)'D_a(\boldsymbol{y}-\hat{\boldsymbol{y}}_J) \\
&= s_J^2+s_C^2, \qquad (11)
\end{aligned}$$

where

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

Theorem 2 *The bias term B_{CV}, when we estimate R_{PE} using the cross-validation criterion* CV*, can be expressed as*

$$\begin{aligned} B_{\mathrm{CV}} &= \mathrm{E}^*_{\boldsymbol{y}}[CV] - R_{\mathrm{PE}} \\ &= \left(\sum_{\alpha=1}^{n} \frac{c_\alpha^2}{1-c_\alpha} \right) \sigma_0^2 + \tilde{\delta}_J^2, \end{aligned}$$

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

$$B_{\mathrm{CV}} = \left(\sum_{\alpha=1}^{n} \frac{c_\alpha^2}{1-c_\alpha} \right) \sigma_0^2. \tag{12}$$

Proof In Theorem 1, it is readily apparent that

$$R_{\mathrm{PE}} = \mathrm{E}[s_J^2] + 2(j+1)\sigma_0^2.$$

On the other side, (11) yields $\mathrm{E}^*_{\boldsymbol{y}}[\mathrm{CV}] = \mathrm{E}^*_{\boldsymbol{y}}[s_J^2] + \mathrm{E}^*_{\boldsymbol{y}}[s_C^2]$. Therefore,

$$B_{\mathrm{CV}} = \mathrm{E}^*_{\boldsymbol{y}}[s_C^2] - 2(j+1)\sigma_0^2.$$

We can write s_C^2 as follows.

$$\begin{aligned} s_C^2 &= \{(I_n - P_J)\boldsymbol{y}\}' D_a \{(I_n - P_J)\boldsymbol{y}\} \\ &= \mathrm{tr}\{(I_n - P_J)\boldsymbol{y}\}' D_a \{(I_n - P_J)\boldsymbol{y}\} \\ &= \mathrm{tr} D_a \{(I_n - P_J)\boldsymbol{y}\}\{(I_n - P_J)\boldsymbol{y}\}' \\ &= \mathrm{tr} D_a \{(I_n - P_J)\{(\boldsymbol{y} - \boldsymbol{\eta}) + \boldsymbol{\eta}\}\{(\boldsymbol{y} - \boldsymbol{\eta}) + \boldsymbol{\eta}\}'(I_n - P_J)\}. \end{aligned}$$

This yields the following reductions, which imply the main result.

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

If the true model is contained in the model M_J, then $(I_n - P_J)\boldsymbol{\eta} = \mathbf{0}$ and hence $\tilde{\delta}^2 = 0$. This completes the proof.

It is natural to assume that $c_\alpha = O(n^{-1})$ because $0 \le c_\alpha < 1$ and $\sum_{\alpha=1}^{n} c_\alpha = k$. Then

$$\sum_{j=\alpha}^{n} \frac{c_\alpha^2}{1-c_\alpha} \le \frac{1}{1-\bar{c}} \sum_{\alpha=1}^{n} c_\alpha^2 = O(n^{-1}),$$

where $\bar{c} = \max_\alpha c_\alpha$. This fact implies that $B_{\mathrm{CV}} = O(n^{-1})$; consequently, CV is asymptotically unbiased when the true model is contained in the candidate model. On the other hand, we have shown that PE is exactly unbiased when the true model is contained in the full model.

4 Multivariate version of PE

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

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

A multivariate linear model is given as

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

where the error terms $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are mutually independent; each 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

$$M_J : Y = X_J \mathcal{B}_J + \mathcal{E},$$

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

$$\begin{aligned} M_* : Y &= (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n)' + (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)' \\ &= \eta + \mathcal{E}, \end{aligned} \tag{13}$$

where the error terms $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are mutually independent. 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 $\boldsymbol{\eta}_\alpha$ under M_J. The measure (4) for goodness of fit of a candidate model M_J is extended as

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

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 (13). In this equation, E_Y^* and E_Z^* respectively denote the expectations with respect Y and Z, where Y and Z are distributed according to the true model (13). Then we can express R_{PE} as

$$\begin{aligned} R_{\mathrm{PE}} &= \sum_{\alpha=1}^{n} \mathrm{E}_Y^* \mathrm{E}_Z^* [(\boldsymbol{\eta}_\alpha - \hat{\boldsymbol{y}}_{\alpha J})'(\boldsymbol{\eta}_\alpha - \hat{y}_{\alpha J})] + n\mathrm{tr}\Sigma_0 \\ &= \mathrm{E}_Y^* [\mathrm{tr}(\eta - \hat{Y})'(\eta - \hat{Y})] + n\mathrm{tr}\Sigma_0. \end{aligned} \tag{14}$$

In a cross-validation for the multivariate prediction error (14), $\boldsymbol{y}_\alpha$ is predicted by the predictor $\hat{\boldsymbol{y}}_{(-\alpha)J}$ based on the data set obtained by removing the αth observation $(\boldsymbol{y}_\alpha, \boldsymbol{x}_\alpha)$; in addition, R_{PE} is estimated by

$$\mathrm{CV} = \sum_{\alpha=1}^{n} (\boldsymbol{y}_\alpha - \hat{\boldsymbol{y}}_{(-\alpha)J})'(\boldsymbol{y}_\alpha - \hat{\boldsymbol{y}}_{(-\alpha)J}).$$

Using the same procedure as that used in the univariate case, we have

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

Now, our interest is an extension of PE to a multivariate case. Let S_J and S_F respectively represent the matrices of sums of squares and products because of errors under M_J and M_F. These matrices are given as

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

where

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

As an estimator of (14), we consider

$$\mathrm{PE} = \mathrm{tr}S_J + \frac{2(j+1)}{n-k-1}\mathrm{tr}S_F. \tag{15}$$

Then the following result, which is an extension of Theorem 3.1, is demonstrated.

Theorem 3 *Presuming that the true model M_* is contained in the full model M_F, then PE in (15) is an unbiased estimator of the multivariate prediction error R_{PE} in (14).*

Proof By an argument similar to that used in the univariate case, we can show that

$$\mathrm{E}_Y^*\mathrm{E}_Z^*[(Z-\hat{Y}_J)'(Z-\hat{Y}_J)] = (n+j+1)\Sigma_0 + \Delta_J,$$
$$\mathrm{E}_Y^*[(Y-\hat{Y}_J)'(Y-\hat{Y}_J)] = (n-j-1)\Sigma_0 + \Delta_J,$$

where $\Delta_J = \eta'(I_n - P_J)\eta$. Furthermore, because 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 MC_p criteria in the univariate case have been extended (Fujikoshi and Satoh 1997) as

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

respectively. For the relative performance of C_p and MC_p for estimators of $\tilde{R}_{\mathrm{PE}}$ in (15), see Fujikoshi and Satoh (1997). □

5 Simulation experiments

In this section, we attempt to give an impression of the relative performances of PE and CV as estimators of their target risk function R_{PE} through simulation experiments. Furthermore, we also examine the relative performances on selection of the true model, more precisely the minimal model which includes the true model or its approximation. The performances are examined for PE and CV, in addition to those of C_p, MC_p, AIC, CAIC, and MAIC, although their target risk functions are not identical. In fact, the risk function for PE and CV is the expected prediction error R_{PE} in (4). For C_p and MC_p it is the expected standardized prediction error given by PE^2/σ_0^2. For AIC, CAIC, and MAIC, it is the expected log-predictive likelihood (8). Ideally a choice of model selection criteria should depend on what the target model is. Furthermore, it depends on assumptions of the models considered. Strictly, the comparison should be done for CV and PE, for C_p and MC_p, or AIC, CAIC, and MAIC. However, in most of the practical data analysis, it is difficult to determine a target risk function clearly. On the other hand, all the seven criteria may be used as a criterion on selection of the minimal model which includes the true model or its approximation, which is called the minimal model simply.

In our simulation experiments we consider the following three cases.

Case (i): The true model is defined by

$$M_*: \ y_\alpha = 2.5 - 10x_\alpha + 10x_\alpha^2 + \varepsilon_\alpha, \quad \alpha = 1, \ldots, n,$$

where $x_\alpha = (\alpha-1)/(n-1)$, $\alpha = 1, \ldots, n$, and the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are mutually independent; each of them has the standard normal distribution $N(0, 1)$. The candidate models considered are polynomial regression models

$$M_k : y_\alpha = \beta_0 + \beta_1 x_\alpha + \cdots + \beta_k x_\alpha^k + \varepsilon_\alpha, \quad \alpha = 1, \ldots, n,$$

for $k = 0, 1, 2, 3$, where the coefficients $\beta_0, \ldots, \beta_k$ are unknown, and the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are mutually independent and have the same mean 0 and the same unknown variance σ^2. The models M_2 and M_3 involve the true model.

Case (2): The true model is the same as the one in case (1) except for that the true mean of $\mathrm{E}(y_\alpha)$ is given by

$$\mathrm{E}(y_\alpha) = 10 \times \exp(0.5 \times x_\alpha).$$

In this case all the candidate models $M_k, k = 0, 1, 2, 3$ do not involve the true model. However, the true mean can be approximated by some polynomial regression model since

$$\exp(0.5 \times x) = 0.5 \times x + (0.5 \times x)^2 + (0.5 \times x)^3 + \cdots .$$

The true mean is nearly a straight line when $0 < x < 1$.

Case (3): The true model is the same as the one in case (1) except for that the error terms $\varepsilon_1, \ldots, \varepsilon_n$ in the true model are mutually independent; each of them has the uniform distribution $U(-\sqrt{3}, \sqrt{3})$. In this case the models M_2 and M_3 include the true model, but under normality assumption of error terms all the models M_k do not include the true model.

Then, 10^5 samples of each sample size were generated from each of the true model. Here, the candidate models are considered only for all hierarchical models M_0, M_1, M_2, M_3. In all the candidate models $M_k, k = 0, 1, 2, 3$ it is not assumed that the distributions of the error terms are normal. However, for use of AIC, CAIC and MAIC, we assume that the distributions of the error terms are normal. Our simulation results are given in Tables 1, 2, 3, 4, 5 and 6 for $n = 10$, $n = 25$ and $n = 50$.

Table 1 Estimation of the risk function R_{PE} in case (1)

n	M_0	M_1	M_2	M_3
10				
R_{PE}	19.05	20.05	13.00	13.99
CV	21.02	27.75	15.68	22.96
PE	19.05	20.08	13.14	14.16
25				
R_{PE}	42.13	43.11	27.98	29.00
CV	43.64	47.53	28.65	30.47
PE	42.22	43.24	28.03	29.03
50				
R_{PE}	81.01	82.02	53.05	54.01
CV	82.30	85.53	53.33	54.63
PE	81.04	82.03	53.06	54.06

Table 2 Relative frequencies selected by seven criteria in case (1)

n	M_0	M_1	M_2	M_3
10				
CV	0.21	0.01	0.61	0.16
PE	0.16	0.01	0.64	0.19
C_p	0.16	0.01	0.64	0.19
MC_p	0.28	0.01	0.59	0.12
AIC	0.11	0.01	0.61	0.27
CAIC	0.61	0.01	0.37	0.01
MAIC	0.69	0.01	0.29	0.00
25				
CV	0.02	0.00	0.79	0.19
PE	0.01	0.00	0.81	0.17
C_p	0.01	0.00	0.81	0.17
MC_p	0.02	0.00	0.83	0.15
AIC	0.01	0.00	0.78	0.20
CAIC	0.03	0.00	0.86	0.11
MAIC	0.04	0.00	0.88	0.08
50				
CV	0.00	0.00	0.83	0.17
PE	0.00	0.00	0.84	0.16
C_p	0.00	0.00	0.84	0.16
MC_p	0.00	0.00	0.84	0.16
AIC	0.00	0.00	0.82	0.18
CAIC	0.00	0.00	0.86	0.14
MAIC	0.00	0.00	0.88	0.12

Table 3 Estimation of the risk function R_{PE} in case (2)

n	M_0	M_1	M_2	M_3
10				
R_{PE}	53.46	12.06	12.88	13.90
CV	63.98	13.05	15.49	22.54
PE	53.82	12.19	12.96	13.94
25				
R_{PE}	53.83	12.25	13.03	14.04
CV	63.90	13.02	15.42	22.46
PE	53.75	12.17	12.94	13.94
50				
R_{PE}	233.13	52.76	52.96	53.94
CV	239.98	52.99	53.31	54.61
PE	232.48	52.81	53.04	54.05

Table 4 Relative frequencies selected by seven criteria in case (2)

n	M_0	M_1	M_2	M_3
10				
CV	0.00	0.68	0.21	0.10
PE	0.00	0.68	0.18	0.15
C_p	0.00	0.68	0.18	0.15
MC_p	0.00	0.79	0.13	0.08
AIC	0.00	0.58	0.21	0.21
CAIC	0.00	0.93	0.06	0.01
MAIC	0.00	0.97	0.03	0.00
25				
CV	0.00	0.68	0.21	0.11
PE	0.00	0.69	0.20	0.11
C_p	0.00	0.69	0.20	0.11
MC_p	0.00	0.72	0.19	0.09
AIC	0.00	0.66	0.22	0.13
CAIC	0.00	0.78	0.17	0.05
$MAIC$	0.00	0.80	0.15	0.04
50				
CV	0.00	0.63	0.26	0.10
PE	0.00	0.64	0.26	0.10
C_p	0.00	0.64	0.26	0.10
MC_p	0.00	0.65	0.26	0.09
AIC	0.00	0.62	0.27	0.11
CAIC	0.00	0.68	0.25	0.08
MAIC	0.00	0.69	0.24	0.07

Table 5 Estimation of the risk function R_{PE} in case (3)

n	M_0	M_1	M_2	M_3
10				
R_{PE}	19.05	20.05	12.98	13.96
CV	21.05	27.77	15.62	22.69
PE	19.07	20.08	13.10	14.10
25				
R_{PE}	42.17	43.17	28.03	29.06
CV	43.66	47.51	28.59	30.36
PE	42.23	43.22	27.96	28.92
50				
R_{PE}	81.17	82.19	52.92	53.90
CV	82.34	85.58	53.34	54.64
PE	81.09	82.07	53.06	54.07

Table 6 Relative frequencies selected by seven criteria in case (3)

n	M_0	M_1	M_2	M_3
10				
CV	0.22	0.01	0.61	0.16
PE	0.16	0.01	0.64	0.19
C_p	0.16	0.01	0.64	0.19
MC_p	0.29	0.01	0.58	0.12
AIC	0.11	0.00	0.61	0.28
CAIC	0.66	0.01	0.32	0.02
MAIC	0.73	0.01	0.25	0.01
25				
CV	0.01	0.00	0.79	0.19
PE	0.01	0.00	0.81	0.18
C_p	0.01	0.00	0.81	0.18
MC_p	0.02	0.00	0.82	0.16
AIC	0.01	0.00	0.78	0.21
CAIC	0.03	0.00	0.86	0.12
MAIC	0.03	0.00	0.88	0.09
50				
CV	0.00	0.00	0.83	0.17
PE	0.00	0.00	0.84	0.16
C_p	0.00	0.00	0.84	0.16
MC_p	0.00	0.00	0.85	0.15
AIC	0.00	0.00	0.82	0.18
CAIC	0.00	0.00	0.87	0.13
MAIC	0.00	0.00	0.88	0.12

The main aim is to find the target model which has minimal risk, or the minimal model. The target model is usually equal to the minimal model when the sample size is large. However, these two models are not the same when the sample size is small, as being seen in our simulation experiments.

Tables 1, 3 and 5 demonstrate that the biases of all PE and CV are reduced as the sample sizes are increasing. For the performances as an estimator of the target risk function R_{PE}, PE is very good for all cases. However, the performances of CV are not so good for small sample sizes $n = 10$ and $n = 25$. In fact, CV has a strong tendency of overestimating its risk for overspecified models. For $n = 50$ or a large sample case, the performance CV becomes good as PE.

The percentages of selecting the minimal model increase concomitant with the increasing sample size and the size of the true model. As pointed out theoretically [see (7)], the selection probabilities of PE and C_p are coincident for all cases. On the other hand, CV displays a tendency of choosing different models in comparison to other models, but PE shows similarity to all of the other criteria except CV.

Table 7 Estimation of the risk function R_A in case (1)

n	M_0	M_1	M_2	M_3
10				
R_A	36.31	37.49	38.97	46.00
AIC	36.89	38.22	31.37	31.54
CAIC	38.60	42.22	39.37	46.54
MAIC	36.90	39.32	38.40	45.36

It may be noted that the differences between the target model and the minimal model are larger in R_A than in $R_{\rm PE}$ or $\tilde{R}_{\rm PE}$, when the sample size is small. For example, the target model of $R_{\rm PE}$ and the minimal model in case (1) with $n = 10$ are the same. On the other hand, the risk R_A and its estimators AIC, CAIC and MAIC are given as follows.

From Table 7 we can see that the target model of R_A is M_0, but the minable model is M_2. Furthermore, AIC underestimates for the overspecified models which include the true model, but the performances of CAIC and MAIC are very good.

From Tables 5 and 6 we can see an effect of nonnormality. It is seen that the effect is smaller in PE, CV, C_p and MC_p than in AIC, CAIC and MAIC.

Through the simulation experiments and the results in Sects. 2 and 3 we can summarize our conclusions on PE as follows. The criteria PE and CV can be regarded as the same target risk function $R_{\rm PE}$. Then, we can say that PE is smaller biases than CV in the estimation of $R_{\rm PE}$. The target risk functions of C_p and PE are different on the point whether the expected prediction error is standardized or not. As the estimators of these target risk functions PE is unbiased under a week assumption, but C_p is not unbiased, though MC_p is unbiased under a stronger assumption. This suggests that PE can be used in a very weak assumption of the distributions of the error terms. The values of PE depend on the variances of data, but the ones of C_p does not depend on them. So, C_p may be used for comparison of a set of data. For the percentages of selecting the minimal model, PE and C_p have the same selection probabilities. In general, PE (CV, C_p, MC_p) is robust than AIC (CAIC, MAIC).

Acknowledgments The authors would like to thank the associated editor and the referee for their valuable comments.

References

Akaike, H. (1973). Informaiton theory and an extension of the maximum likelihood principle. In B. N. Petrov, F. Csáki, (Eds.), *2nd International symposium on information theory* (pp. 267–281). Budapest: Akadémia Kiado.

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

Allen, D. M. (1974). The relationship between variable selection and data augumentation, and a method for prediction. *Technometrics, 16*, 125–127.

Bedrick, E. D., Tsai, C. L. (1994). Model selection for multivariate regression in small samples. *Biometrics*, *50*, 226–231.

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

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

Haga, Y., Takeuchi, K., Okuno, C. (1973). New criteria for selecting of variables in regression model. *Quality (Hinshitsu, Journal of the Japanese Society for Quality Control)*, *6*, 73–78 (in Japanese).

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

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

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

Stone, M. (1974). Cross-validatory choice and assesment of statistical predictions (with Discussion). *Journal of the Royal Statistical Society, B, 36*, 111–147.

Sugiura, N. (1978). Futher analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics: Theory and Methods*, *7*, 13–26.