THE VARIABLE SELECTION PROBLEM FOR PREDICTION ON THE PREDICTION AREA DIFFERENT FROM SAMPLE AREA

Takakatsu Inoue*

In the regression model \( Y_a = x_a' \beta + \epsilon_a \) \((a=1, 2, \ldots, n)\), we discuss the variable selection problem for prediction of the future dependent variable \( Y^* \) on the future sample point \( x^* \), where \( x^* \) is assumed to be a vector of random variables having a known probability density function. As the risk function for variable selection, the expectation with respect to \( x^* \) of Prediction Mean Square Error is used. Two criteria for variable selection are proposed and discussed on the basis of numerical examples.

1. Introduction

In regression analysis, several information criteria for variable selection (or model selection) including AIC (Akaike's Information Criterion [1]), \( C_p \) (Mallows' \( C_p \) statistic [2]) etc. were proposed as the methods of variable selection, and have been examined by various authors from the different points of view (for example see Shibata [6], Hashimoto et. al. [3] and Nishii [5]). These criteria are considered to have been proposed for predicting the future dependent variable \( Y^* \) on the sample points \( x_a \) \((a=1, 2, \ldots, n)\).

In this paper, we consider the above problem in situation where the prediction area, on which we predict the future dependent variable \( Y^* \), is different from the sample area on which \( y_a \) \((a=1, 2, \ldots, n)\) are observed. We propose two criteria for this model selection and discuss them. As a risk function for the model selection, we use the expectation of prediction mean square error with respect to the future independent variable \( x^* \).

By the way, Allen [2] proposed a criterion for predicting \( Y^* \) on a particular future sample point \( x^* \). Then, for multiple future sample points, the process of variable selection is repeated for each point and different models may be selected for each future sample point (see Thompson [7]). Thus, it is desirable to employ an efficient algorithm in performing the necessary computations. The criteria proposed in this paper include Allen's criterion (see section 3).

2. Formulation

Let \( Y \) and \( x_q=(x_0, x_1, \ldots, x_{q-1})' \) be a dependent univariate random variable and a vector of independent \( q \)-variate variables respectively which are considered to be available for explaining the fluctuations of \( Y \). Given the sample \( \{(x_{qa}, Y_a); a=1, 2, \ldots, n\} \), let us assume the following multiple linear regression model;

\[
(2.1) \quad Y_a = x_{qa}' \beta_a + \epsilon_a \quad (a=1, 2, \ldots, n),
\]

or in matrix notation

---


* Faculty of Engineering, Chiba University, 1-33, Yayoi-cho, Chiba 260 Japan
(2.2) \[ Y = Xq \beta_q + \varepsilon_q, \quad (n \times 1) \quad (n \times q) \quad (q \times 1) \quad (n \times 1) \]

where $\beta_q$ is the unknown regression coefficients vector, the rank of design matrix $X_q$ is assumed to be $q$, $x_{\alpha} = 1$ for $\alpha = 1, 2, \ldots, n$ and $\varepsilon_q$ be a vector of random variables having expected value 0 and dispersion matrix $\sigma^2 I_n$, i.e., $\varepsilon_q \sim (0, \sigma^2 I_n)$.

Let us consider the problem of predicting the value of the future dependent variable $Y^*$ on the future sample point $x_q^*$, where $x_q^*$ is assumed to be a vector of $q$-variates random variables having a known distribution. So we consider the case in which future sample points may be different from sample points $x_{\alpha}$ ($\alpha = 1, 2, \ldots, n$). Now "sample area" means the region $R = \{x_{\alpha}; \alpha = 1, 2, \ldots, n\}$ and "prediction area" means the region $R^* = \{x_q^*\}$ on which we will predict the future dependent variable $Y^*$. Further let us assume that $Y^*$ on any fixed $x_q^*$ in $R^*$ can be expressed as

(2.3) \[ Y^* = x_q^* \beta_0 + \varepsilon_q^*, \quad \varepsilon_q^* \sim (0, \sigma^2), \]

where $\varepsilon_q^*$ is assumed to be independent of $\varepsilon_q$ of (2.2).

By the way, this regression model may include independent variables which are of no use for the prediction of $Y^*$ on $R^*$. For example, variables whose regression coefficients are zero or nearly equal to zero should be removed from the regression model (2.2) and (2.3). For this reason we have a need of model selection.

There are $q-1 C_{p-1}$ models when $(p-1)$ variables from $(q-1)$ variables are included in the model, where the variable $x_0$ is assumed to be included in any model. With each model we associate the suffix vector $i(p) = (i_0, i_1, \ldots, i_r, \ldots, i_{q-1})$, where $i_r = 1$ if the variable $x_r$ is included in the model and $i_r = 0$, otherwise, and $p$ is the number of the independent variables included in the model. We express the regression model with the suffix vector $i(p)$ by

(2.4) \[ Y = X(i(p)) \beta_{i(p)} + \varepsilon_{i(p)}, \quad \text{or} \quad Y = X_i \beta_i + \varepsilon_i \]

and denotes it by $M[i(p)]$. For example, in the case where $q = 5$ and the suffix vector is $5(2) = (1, 0, 1, 0, 0)$, $M[5(2)]$ represents the model with the independent variables $x_0$, $x_1$. The model $M[i(p)]$ with the suffix vector $i(p) = (i_0, i_1, \ldots, i_r, \ldots, i_{q-1})$ is defined to be included in the model $M[j(p)]$ with $j(p) = (j_0, j_1, \ldots, j_r, \ldots, j_{q-1})$ and denoted by

(2.5) \[ M[i(p)] \subseteq M[j(p)], \]

if $i_r \leq j_r$ is satisfied for $r = 0, 1, \ldots, q-1$. And we define $M[i(p)] = M[j(p)]$ when $i_r = j_r$ for all $r$. And the model $M[k(p)]$ with the suffix vector $k(p) = (k_0, k_1, \ldots, k_r, \ldots, k_{q-1})$ is defined as

(2.6) \[ M[k(p)] = M[i(p)] \cup M[j(p)], \]

if $k_r = \max(i_r, j_r)$.

For convenience, independent variables which have non-zero regression coefficients of (2.2) and (2.3) are assumed to be $x_i = (x_0, x_1, \ldots, x_{q-1})$ and "true regression model"'s corresponding to (2.2) and (2.3) are assumed to be expressed as follows.

(2.7) \[ Y = X_i \beta_i + \varepsilon_i, \quad \varepsilon_i \sim (0, \sigma^2 I_n) \]

(2.8) \[ Y^* = x_q^* \beta_0 + \varepsilon_q^*, \quad \varepsilon_q^* \sim (0, \sigma^2) \]

And we denote the models (2.2), (2.7) by $M[q]$, $M[i]$, respectively.
3. Risk Function

In this section, we consider a risk function for the model selection based on $E_{x*Y} \{(Y^*-x_i^* \beta_i)^2\}$, where $E_{x*Y}$ means the expectation with respect to future variables $Y^*$ and $x^*$ and $\beta_i$ is a regression coefficient vector of the model $M[i(p)]$. For any fixed model $M[i(p)]$, the above risk function attains the minimum value under the following relation

$$\beta_i = \Delta t^{-1} \Delta t \beta_i,$$

where $\Delta t = E(x_i^* x_i^*)$, $\Delta t = E(x_i^* x_i^*)$. So, $\beta_i$ is now uniquely defined by the normal equation (3.1) in population. It is necessary to estimate $\beta_i$ based on the sample $\{(x_a, Y_a); a=1, 2, \ldots, n\}$, because the value of $\beta_i$ (and so $\beta_i$) is unknown. Then the final risk function should be given by $E_{x*Y} E_r \{(Y^*-x_i^* \beta_i)^2\}$, where $\tilde{\beta}_i$ is an estimator of $\beta_i$ and $E_r$ means the expectation with respect to $Y$. Using the equation (3.1), the final risk function is expressed as

$$r(Y^*, \tilde{\beta}_i) = E_{x*Y} E_r \{(Y^*-x_i^* \beta_i)^2\}$$

$$= E_{x*Y} \{(Y^*-x_i^* \beta_i)^2\} + E_{x*Y} \{(x_i^* \beta_i - x_i^* \beta_i)^2\}$$

$$= \sigma_i^2 + (\beta_i' \Delta u \beta_i - \beta_i' \Delta u \beta_i)$$

$$+ \text{tr}[\Delta u E_r \{(\beta_i - \tilde{\beta}_i)(\beta_i - \tilde{\beta}_i)'\}].$$

Now let us consider the case in which the ordinary least square estimator (O.L.S.E.) $\hat{\beta}_i = (X'X_i)^{-1}X_i'Y$ is used as the estimator of regression coefficients vector in the model $M[i(p)]$. In this case, we have

$$E_r \{(\beta_i - \hat{\beta}_i)(\beta_i - \tilde{\beta}_i)'\}$$

$$= (\beta_i - \beta(X)_i)(\beta_i - \beta(X)_i)' + \sigma_i^2(X_i'X_i)^{-1},$$

where

$$\beta(X)_i = (X_i'X_i)^{-1}X_i'Y.$$

Thus the risk function $r(Y^*, \hat{\beta}_i)$ is expressed as follows:

$$r(Y^*, \hat{\beta}_i) = \sigma_i^2 + (\beta_i' \Delta u \beta_i - \beta_i' \Delta u \beta_i)$$

$$+ (\beta_i - \beta(X)_i)' \Delta u (\beta_i - \beta(X)_i)$$

$$+ \sigma_i^2 \text{tr}[\Delta u (X_i'X_i)^{-1}].$$

Now each term of the right-side of (3.5) can be taken as follows:

- the first term: the variance of $Y^*$ around the true regression plane $x_i^* \beta_i$,
- the second term: the model bias of the model $M[i(p)]$ on the prediction area,
- the third term: the estimation bias of $\hat{\beta}_i$ which vanishes if $E_r(\hat{\beta}_i) = \beta_i$, or $M[i(p)] \equiv M[t]$, the fourth term: the estimation variance of $\hat{\beta}_i$ on the prediction area.

Under the condition that

$$\Delta u = n^{-1}X_q'X_q,$$

the risk function $E_{x*Y} E_r \{(Y^*-x_i^* \hat{\beta}_i)^2\}$, denoted by $r_c(Y^*, \hat{\beta}_i)$, is given by

$$r_c(Y^*, \hat{\beta}_i) = \sigma_i^2 + (\beta_i' (n^{-1}X_i'X_i) \beta_i - \beta(X)_i'(n^{-1}X_i'X_i) \beta(X)_i) + n^{-1} \sigma_i^2 \beta_i.$$
By using the projection matrix \( Q(X) = X(X'X)^{-1}X' \), the second term of (3.7) is expressed as

\[
(3.8) \quad n^{-1} \beta_1' \sigma^2 \epsilon \{ I_n - Q(X) \} X_1 \beta_1.
\]

Note that the condition (3.6) holds in the case of the prediction only on the sample area; the future sample point \( x_{q*} \) is implicitly assumed to have a discrete uniform distribution on the sample area. Further, if \( x_{q*} = x_{q*} x_{q*}' \) (\( x_{q*} \): the particular future sample point), the risk function (3.5) is equivalent to Allen's criterion [2]. The second term, the third term of (3.5) and the second term of (3.7) vanish in any model \( M[i(p)] \) which satisfies the relation \( M[i(p)] \equiv M[t] \). Note that the fourth term of (3.5) and the third term of (3.7) are of order \( O(n^{-1}) \) and the other terms of (3.5) and (3.7) are of order \( O(n^0) \). From (3.5) and (3.7), the optimum model based on (3.5) and that based on (3.7) are different in general.

By the way, if \( M[q] \equiv M[t] \), the following statistic

\[
(3.9) \quad l_c(\hat{\beta}_t) = n^{-1} \{ (Y - X \hat{\beta}_t)'(Y - X \hat{\beta}_t) + (Y - X \hat{\beta}_t)'(Y - X \hat{\beta}_t) + 2p/(n - q) \}
\]

is an unbiased estimator of (3.7), so \( l_c(\hat{\beta}_t) \) can be used as a criterion for model selection on the sample area. The estimator \( l_c(\hat{\beta}_t) \) is essentially equal to \( C_p \) statistic of Mallows [3]. If an unbiased estimator of (3.5) is obtained, it can be used as a criterion for model selection on the prediction area. Note that the estimator \( \hat{\beta}_t \) of \( \beta_t \) has the estimation bias of order \( O(n^0) \) in addition to the estimation variance of order \( O(n^{-1}) \) because \( E r(\hat{\beta}_t) \neq \beta_t \).

In the next section, let us consider an unbiased estimator of \( \beta_t \), and introduce two criteria for prediction on the prediction area.

4. Criteria for Prediction

Under the assumption that \( \Delta_W \) is a positive definite matrix of order \( q \), let us consider the estimator of \( \beta_t \) such as

\[
(4.1) \quad (X'WWX_t)^{-1}(X'WY),
\]

where \( W \) is assumed to be a positive definite matrix of order \( n \). First we construct a matrix \( W \) satisfying the unbiased condition

\[
(4.2) \quad E r((X'WWX_t)^{-1}(X'WY)) = \beta_t \quad \text{for any } M[i(p)].
\]

From that

\[
(4.3) \quad E r((X'WWX_t)^{-1}(X'WY)) = (X'WWX_t)^{-1}(X'WWX_t) \beta_t
\]

and that \( \beta_t = \Delta_W^{-1} \Delta_W \beta_t \), we can get an unbiased estimator of \( \beta_t \) if there exists the matrix \( W \) satisfying the following equations:

\[
(4.4) \quad X'WWX_t = \Delta_W, \quad X'WX_t = \Delta_W.
\]

The equations of (4.4) holds if

\[
(4.5) \quad X'WX_t = \Delta_W,
\]

where \( X_t \) is the design matrix corresponding to \( M[j(r)] \) such that

\[
(4.6) \quad M[j(r)] = M[i(p)] \cup M[t].
\]
THE VARIABLE SELECTION PROBLEM FOR PREDICTION

(4.6) satisfies the relations \( M[j(r)] \subseteq M[i(p)] \subseteq M[q] \), \( M[i] \subseteq M[j(r)] \subseteq M[q] \) and therefore \( p \leq r \leq q \), \( t \leq r \leq q \). Further, the equation (4.5) holds if

\[
X_q'WX_q = I_{n-q}.
\]

where \( X_q \) is the design matrix corresponding to \( M[q] \). (Note that all elements of \( X_r'WX_r \) exist in both \( X_j'WX_j \) and \( X_q'WX_q \), and that all elements of \( I_{n-q} \) exist in \( I_{n-j} \).)

A matrix \( W_{ji} \) which satisfies the equation (4.5) is given by

\[
W_{ji} = X_j(X_j'X_j)^{-1}\Lambda_{ji}(X_j'X_j)^{-1}X_j',
\]

(see Appendix A). Similarly, a matrix \( W_q \) which satisfies (4.5)' is obtained as

\[
W_q = X_q(X_q'X_q)^{-1}\Lambda_{qi}(X_q'X_q)^{-1}X_q'.
\]

By substituting \( W_{ji} \) of (4.7) and \( W_q \) of (4.7)' for \( W \) of (4.1) respectively, we obtain two unbiased estimators of \( \beta_i \);

\[
\hat{\beta}_i = (X_i'W_{ji}X_i)^{-1}X_i'W_{ji}Y,
\]

\[
\tilde{\beta}_i = (X_i'W_qX_i)^{-1}X_i'W_qY.
\]

These estimators satisfy the following relations,

\[
E(\hat{\beta}_i) = \beta_i, \quad V(\hat{\beta}_i) = \sigma^2(I + \Lambda_i)X_i'W_{ji}X_i\Lambda_i^{-1} \quad \text{for any } M[i(p)].
\]

\[
E(\tilde{\beta}_i) = \beta_i, \quad V(\tilde{\beta}_i) = \sigma^2(I + \Lambda_i)X_i'W_qX_i\Lambda_i^{-1} \quad \text{for any } M[i(p)].
\]

Further, the risk function \( r(Y^*, \beta_i) \) and \( r(Y^*, \tilde{\beta}_i) \) are expressed as follows.

\[
r(Y^*, \hat{\beta}_i) = \sigma^2 + tr(W_{ji}X_i\Lambda_i^{-1}X_i'W_{ji}X_iX_i\Lambda_i^{-1}X_i'),
\]

\[
r(Y^*, \tilde{\beta}_i) = \sigma^2 + tr(W_qX_i\Lambda_i^{-1}X_i'W_qX_iX_i\Lambda_i^{-1}X_i'),
\]

where \( Z_i = W_{ji}X_i X_i, Z_i = W_qX_i X_i \) in (4.10), \( Z_i = W_{ji}X_i X_i, Z_i = W_qX_i X_i \) in (4.10)' and \( Q(Z_i) = Z_i(Z_i'Z_i)^{-1}Z_i' \) in (4.10) and (4.10)' (see Appendix A).

An unbiased estimator of (4.10) and that of (4.10)' (denoted by \( l(\hat{\beta}_i) \) and \( l(\tilde{\beta}_i) \), respectively) are expressed as follows (see Appendix B);

\[
l(\hat{\beta}_i) = (Y - X\hat{\beta}_i)'W_{ji}(Y - X\hat{\beta}_i) + \frac{1}{2} tr(W_{ji}X_i\Lambda_i^{-1}X_i')(n - q),
\]

\[
l(\tilde{\beta}_i) = (Y - X\tilde{\beta}_i)'W_q(Y - X\tilde{\beta}_i) + \frac{1}{2} tr(W_qX_i\Lambda_i^{-1}X_i')(n - q).
\]

We can do model selection for prediction on the prediction area \( R^* \) by using \( l(\hat{\beta}_i) \) or \( l(\tilde{\beta}_i) \). Note that the criterion based on \( l(\tilde{\beta}_i) \) can not be used directly in our situation because the model \( M[j(r)] \) of (4.6) is unknown. From the viewpoint of minimizing \( r(Y^*, \beta_i) \), we have a need of model selection even in situation where the model \( M[i] \) is known and the value of \( \beta_i \) is unknown (for example, see EX-1 in section 5). In this situation, the criterion based on \( l(\tilde{\beta}_i) \) is applicable. But this situation is rarely in practice. So, the problem to
make the criterion \( l(\hat{\beta}) \) applicable to practical situation is discussed in section 7.

In the next section, let us show numerical examples.

5. Examples

As examples we consider the following polynomial regression models.

1) the model \( M[q] \)

\[
y_a = \beta_0 + \beta_1 x_a + \beta_2 x_a^2 + \beta_3 x_a^3 + \epsilon_t
\]

\[= 0 + 1.5 x_a + (0.1) x_a^2 + 0 + 0 + \epsilon_t, \quad \epsilon_t \sim (0, \sigma_t^2), \]

where \( \sigma_t^2 = 9.0 \).

2) model \( M[i(p)] \)

- \( M[1(1)]: y_a = \beta_{10} + \epsilon_{1a} \)
- \( M[2(2)]: y_a = \beta_{20} + \beta_{21} x_a + \epsilon_{2a} \)
- \( M[3(3)]: y_a = \beta_{30} + \beta_{31} x_a + \beta_{32} x_a^2 + \epsilon_{3a} \)
- \( M[4(4)]: y_a = \beta_{40} + \beta_{41} x_a + \beta_{42} x_a^2 + \beta_{43} x_a^3 + \epsilon_{4a} \)
- \( M[5(5)]: y_a = \beta_{50} + \beta_{51} x_a + \beta_{52} x_a^2 + \beta_{53} x_a^3 + \beta_{54} x_a^4 + \epsilon_{5a} \)

3) the sample area \( R = \{ x_a: a = 1, 2, \ldots, n \} \)

The sample point \( x_a \) is given by uniformly distributed random digit on \([0, 10]\), sample size \( n \) is taken as \( n=20 \) in EX-1, \( n=50 \) in EX-2, EX-3.

4) prediction area \( R^* = \{ x^*: a \leq x^* \leq b \} \) (the intervals: \([5, 10]\) in EX-1, EX-2 and \([0, 15]\) in EX-3),

distribution: uniform distribution on \( R^* \).

5) true regression model

\[
y^* = \beta_0 + \beta_1 x^* + \beta_2 x^* \epsilon^* + \epsilon^*
\]

\[= 0 + 1.5 x^* + (0.1) x^* + \epsilon^*, \quad \epsilon^* \sim (0, \sigma_t^2). \]

The calculated values of the risk functions of (3.5), (3.7), (4.10) and (4.10)' for the model \( M[i(p)] \) are summarized in the Table 1-a, 2-a and 3-a. And the relative precision in M.S.E.

<table>
<thead>
<tr>
<th>Model</th>
<th>Model-Bias</th>
<th>Est.-Bias</th>
<th>Est.-Var.</th>
<th>( r(Y^*, \hat{\beta}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>M[1(1)]</td>
<td>0.0347</td>
<td>0.7076</td>
<td>0.4500</td>
<td>10.192 (1)</td>
</tr>
<tr>
<td>M[2(2)]</td>
<td>0.0347</td>
<td>0.5068</td>
<td>0.8789</td>
<td>10.420 (2)</td>
</tr>
<tr>
<td>M[3(3)]</td>
<td>0.0</td>
<td>0.0</td>
<td>1.5622</td>
<td>10.562 (3)</td>
</tr>
<tr>
<td>M[4(4)]</td>
<td>0.0</td>
<td>0.0</td>
<td>1.9445</td>
<td>10.945 (4)</td>
</tr>
<tr>
<td>M[5(5)]</td>
<td>0.0</td>
<td>0.0</td>
<td>2.6169</td>
<td>11.617 (5)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Model-Bias</th>
<th>Est.-Bias</th>
<th>Est.-Var.</th>
<th>( r_a(Y^*, \hat{\beta}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>M[1(1)]</td>
<td>1.8676</td>
<td>0.0</td>
<td>0.4500</td>
<td>11.318 (5)</td>
</tr>
<tr>
<td>M[2(2)]</td>
<td>0.3566</td>
<td>0.0</td>
<td>0.9000</td>
<td>10.257 (1)</td>
</tr>
<tr>
<td>M[3(3)]</td>
<td>0.0</td>
<td>0.0</td>
<td>1.3500</td>
<td>10.350 (2)</td>
</tr>
<tr>
<td>M[4(4)]</td>
<td>0.0</td>
<td>0.0</td>
<td>1.8000</td>
<td>10.800 (3)</td>
</tr>
<tr>
<td>M[5(5)]</td>
<td>0.0</td>
<td>0.0</td>
<td>2.2500</td>
<td>11.250 (4)</td>
</tr>
</tbody>
</table>
### Table 1-a. continued

<table>
<thead>
<tr>
<th>Model</th>
<th>Model-Bias</th>
<th>Est.-Bias</th>
<th>Est.-Var.</th>
<th>$r(Y^*, \hat{\beta}_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M[1(1)]</td>
<td>0.0347</td>
<td>0.0</td>
<td>0.7561</td>
<td>9.791 (1)</td>
</tr>
<tr>
<td>M[2(2)]</td>
<td>0.0347</td>
<td>0.0</td>
<td>1.5184</td>
<td>10.553 (2)</td>
</tr>
<tr>
<td>M[3(3)]</td>
<td>0.0</td>
<td>0.0</td>
<td>1.5622</td>
<td>10.562 (3)</td>
</tr>
<tr>
<td>M[4(4)]</td>
<td>0.0</td>
<td>0.0</td>
<td>1.9445</td>
<td>10.945 (4)</td>
</tr>
<tr>
<td>M[5(5)]</td>
<td>0.0</td>
<td>0.0</td>
<td>2.6169</td>
<td>11.617 (5)</td>
</tr>
</tbody>
</table>

### Table 1-b. Relative precision of $\tilde{\beta}_i$, $\tilde{\beta}_i$ to $\hat{\beta}_i$.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\tilde{\beta}_i$</th>
<th>$\tilde{\beta}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M[1(1)]</td>
<td>0.6531</td>
<td>0.6659</td>
</tr>
<tr>
<td>M[2(2)]</td>
<td>1.0958</td>
<td>1.3437</td>
</tr>
<tr>
<td>M[3(3)]</td>
<td>1.0</td>
<td>1.5442</td>
</tr>
<tr>
<td>M[4(4)]</td>
<td>1.0</td>
<td>1.3441</td>
</tr>
<tr>
<td>M[5(5)]</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

### Table 2-a. Behaviour of risk functions in EX-2.

$R = \{x; 0 \leq x \leq 10\}$, $n = 50$, $R^* = \{x^*; 5 \leq x^* \leq 10\}$

<table>
<thead>
<tr>
<th>Model</th>
<th>Model-Bias</th>
<th>Est.-Bias</th>
<th>Est.-Var.</th>
<th>$r(Y^*, \hat{\beta}_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M[1(1)]</td>
<td>0.0347</td>
<td>1.1363</td>
<td>0.1800</td>
<td>10.351 (5)</td>
</tr>
<tr>
<td>M[2(2)]</td>
<td>0.0347</td>
<td>0.5897</td>
<td>0.3564</td>
<td>9.581 (3)</td>
</tr>
<tr>
<td>M[3(3)]</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5969</td>
<td>9.597 (1)</td>
</tr>
<tr>
<td>M[4(4)]</td>
<td>0.0</td>
<td>0.0</td>
<td>0.8470</td>
<td>9.847 (2)</td>
</tr>
<tr>
<td>M[5(5)]</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0955</td>
<td>10.096 (4)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Model-Bias</th>
<th>Est.-Bias</th>
<th>Est.-Var.</th>
<th>$r_d(Y^*, \hat{\beta}_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M[1(1)]</td>
<td>2.5213</td>
<td>0.0</td>
<td>0.1800</td>
<td>11.701 (5)</td>
</tr>
<tr>
<td>M[2(2)]</td>
<td>0.4674</td>
<td>0.0</td>
<td>0.3600</td>
<td>9.827 (3)</td>
</tr>
<tr>
<td>M[3(3)]</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5400</td>
<td>9.540 (1)</td>
</tr>
<tr>
<td>M[4(4)]</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7200</td>
<td>9.720 (2)</td>
</tr>
<tr>
<td>M[5(5)]</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9000</td>
<td>9.900 (4)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Model-Bias</th>
<th>Est.-Bias</th>
<th>Est.-Var.</th>
<th>$r(Y^*, \tilde{\beta}_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M[1(1)]</td>
<td>0.0347</td>
<td>0.0</td>
<td>0.3134</td>
<td>9.348 (1)</td>
</tr>
<tr>
<td>M[2(2)]</td>
<td>0.0347</td>
<td>0.0</td>
<td>0.5835</td>
<td>9.618 (3)</td>
</tr>
<tr>
<td>M[3(3)]</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5969</td>
<td>9.597 (2)</td>
</tr>
<tr>
<td>M[4(4)]</td>
<td>0.0</td>
<td>0.0</td>
<td>0.8470</td>
<td>9.847 (4)</td>
</tr>
<tr>
<td>M[5(5)]</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0955</td>
<td>10.096 (5)</td>
</tr>
</tbody>
</table>
### Table 2-a. Relative precision of $\hat{Y}_i$, $\tilde{Y}_i$ to $\hat{Y}_i$

<table>
<thead>
<tr>
<th>Model</th>
<th>Model-Bias</th>
<th>Est.-Bias</th>
<th>Est.-Var.</th>
<th>$r(Y^*, \tilde{Y}_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M[1(1)]$</td>
<td>0.0347</td>
<td>0.0</td>
<td>0.3298</td>
<td>9.364 (1)</td>
</tr>
<tr>
<td>$M[2(2)]$</td>
<td>0.0347</td>
<td>0.0</td>
<td>0.7260</td>
<td>9.761 (2)</td>
</tr>
<tr>
<td>$M[3(3)]$</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0035</td>
<td>10.037 (3)</td>
</tr>
<tr>
<td>$M[4(4)]$</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0946</td>
<td>10.095 (4)</td>
</tr>
<tr>
<td>$M[5(5)]$</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0955</td>
<td>10.096 (5)</td>
</tr>
</tbody>
</table>

### Table 2-b. Relative precision of $\hat{Y}_i$, $\tilde{Y}_i$ to $\hat{Y}_i$

<table>
<thead>
<tr>
<th>Model</th>
<th>$\hat{Y}_i$</th>
<th>$\tilde{Y}_i$</th>
<th>$\beta_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M[1(1)]$</td>
<td>0.2381</td>
<td>0.2498</td>
<td>$\beta_i$</td>
</tr>
<tr>
<td>$M[2(2)]$</td>
<td>0.6168</td>
<td>0.7674</td>
<td>$\beta_i$</td>
</tr>
<tr>
<td>$M[3(3)]$</td>
<td>1.0</td>
<td>1.7373</td>
<td>$\beta_i$</td>
</tr>
<tr>
<td>$M[4(4)]$</td>
<td>1.0</td>
<td>1.2928</td>
<td>$\beta_i$</td>
</tr>
<tr>
<td>$M[5(5)]$</td>
<td>1.0</td>
<td>1.0</td>
<td>$\beta_i$</td>
</tr>
</tbody>
</table>

### Table 3-a. Behaviour of risk functions in EX-3.

$R = \{ x; 0 \leq x \leq 10 \}$, $n=50$. $R^* = \{ x^*; 0 \leq x^* \leq 15 \}$

<table>
<thead>
<tr>
<th>Model</th>
<th>Model-Bias</th>
<th>Est.-Bias</th>
<th>Est.-Var.</th>
<th>$r(Y^*, \hat{Y}_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M[1(1)]$</td>
<td>2.8125</td>
<td>0.3608</td>
<td>0.1800</td>
<td>12.353 (1)</td>
</tr>
<tr>
<td>$M[2(2)]$</td>
<td>2.8125</td>
<td>8.3837</td>
<td>0.7571</td>
<td>20.958 (3)</td>
</tr>
<tr>
<td>$M[3(3)]$</td>
<td>0.0</td>
<td>0.0</td>
<td>5.0710</td>
<td>14.071 (2)</td>
</tr>
<tr>
<td>$M[4(4)]$</td>
<td>0.0</td>
<td>0.0</td>
<td>43.9480</td>
<td>52.948 (4)</td>
</tr>
<tr>
<td>$M[5(5)]$</td>
<td>0.0</td>
<td>0.0</td>
<td>457.7100</td>
<td>466.710 (5)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Model-Bias</th>
<th>Est.-Bias</th>
<th>Est.-Var.</th>
<th>$r_6(Y^*, \hat{Y}_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M[1(1)]$</td>
<td>2.5213</td>
<td>0.0</td>
<td>0.1800</td>
<td>11.701 (5)</td>
</tr>
<tr>
<td>$M[2(2)]$</td>
<td>0.4674</td>
<td>0.0</td>
<td>0.3600</td>
<td>9.827 (3)</td>
</tr>
<tr>
<td>$M[3(3)]$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5400</td>
<td>9.540 (1)</td>
</tr>
<tr>
<td>$M[4(4)]$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7200</td>
<td>9.720 (2)</td>
</tr>
<tr>
<td>$M[5(5)]$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9000</td>
<td>9.900 (4)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Model-Bias</th>
<th>Est.-Bias</th>
<th>Est.-Var.</th>
<th>$r(Y^*, \tilde{Y}_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M[1(1)]$</td>
<td>2.8125</td>
<td>0.0</td>
<td>1.5562</td>
<td>13.369 (1)</td>
</tr>
<tr>
<td>$M[2(2)]$</td>
<td>2.8125</td>
<td>0.0</td>
<td>3.9879</td>
<td>15.800 (3)</td>
</tr>
<tr>
<td>$M[3(3)]$</td>
<td>0.0</td>
<td>0.0</td>
<td>5.0710</td>
<td>14.071 (2)</td>
</tr>
<tr>
<td>$M[4(4)]$</td>
<td>0.0</td>
<td>0.0</td>
<td>43.9480</td>
<td>52.948 (4)</td>
</tr>
<tr>
<td>$M[5(5)]$</td>
<td>0.0</td>
<td>0.0</td>
<td>457.7100</td>
<td>466.710 (5)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Model-Bias</th>
<th>Est.-Bias</th>
<th>Est.-Var.</th>
<th>$r(Y^*, \tilde{Y}_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M[1(1)]$</td>
<td>2.8125</td>
<td>0.0</td>
<td>94.7270</td>
<td>106.540 (1)</td>
</tr>
<tr>
<td>$M[2(2)]$</td>
<td>2.8125</td>
<td>0.0</td>
<td>281.6100</td>
<td>293.423 (2)</td>
</tr>
<tr>
<td>$M[3(3)]$</td>
<td>0.0</td>
<td>0.0</td>
<td>408.8000</td>
<td>418.800 (3)</td>
</tr>
<tr>
<td>$M[4(4)]$</td>
<td>0.0</td>
<td>0.0</td>
<td>452.1200</td>
<td>461.120 (4)</td>
</tr>
<tr>
<td>$M[5(5)]$</td>
<td>0.0</td>
<td>0.0</td>
<td>457.7100</td>
<td>466.710 (5)</td>
</tr>
</tbody>
</table>
Table 3-b. Relative precision of $\hat{\beta}_i$, $\tilde{\beta}_i$ to $\beta_i$.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\hat{\beta}_i$</th>
<th>$\tilde{\beta}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M[1(1)]</td>
<td>2.8774</td>
<td>175.1512</td>
</tr>
<tr>
<td>M[2(2)]</td>
<td>0.4360</td>
<td>30.7911</td>
</tr>
<tr>
<td>M[3(3)]</td>
<td>1.00</td>
<td>80.8125</td>
</tr>
<tr>
<td>M[4(4)]</td>
<td>1.00</td>
<td>10.2876</td>
</tr>
<tr>
<td>M[5(5)]</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

of $\hat{\beta}_i$ to $\hat{\beta}_i$: M.S.E.$(\hat{\beta}_i)/$M.S.E.$(\tilde{\beta}_i)$ (M.S.E.$(\hat{\beta}_i) = \text{tr}[\Delta_iE_r((\beta_i - \hat{\beta}_i)(\beta_i - \tilde{\beta}_i)^\prime)])$ and that of $\tilde{\beta}_i$ to $\hat{\beta}_i$ are also summarized in the Table 1-b, 2-b and 3-b. Tables 1-a, 2-a, and 3-a show that the third term of (3.5) was larger than the forth term of (3.5) in models M[1(1)] and M[2(2)] such that $M[i(p)] \subseteq M[t]$ in EX-1, EX-2, EX-3, and show that the optimum model based on $r(Y^*, \hat{\beta}_i)$ was different from that based on $r_e(Y^*, \hat{\beta}_i)$ in EX-1 and EX-3. Therefore the criterion based on $r_e(Y^*, \hat{\beta}_i)$ can not be used in our problem. Tables 1-b, 2-b, and 3-b show that the estimator $\tilde{\beta}_i$ is better than two other estimators in the sense of M.S.E.

6. Discussions

In this section, we compare the risk functions $r(Y^*, \hat{\beta}_i)$ and $r(Y^*, \tilde{\beta}_i)$ with the risk function $r(Y^*, \hat{\beta}_i)$. First of all, the estimators $\hat{\beta}_i$, $\tilde{\beta}_i$, and $\bar{\beta}_i$ have following properties (see Table 4):

Table 4. Comparison of estimators of $\beta_i$.

<table>
<thead>
<tr>
<th>models</th>
<th>estimator</th>
<th>$p$</th>
<th>$r$</th>
<th>$q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_i$</td>
<td>$\hat{\beta}_i$</td>
<td>bias</td>
<td>unbiased</td>
<td>unbiased</td>
</tr>
<tr>
<td>$\beta_i$</td>
<td>$\tilde{\beta}_i$</td>
<td>BLUE</td>
<td>BLUE</td>
<td>unbiased</td>
</tr>
<tr>
<td>$\beta_i$</td>
<td>$\bar{\beta}_i$</td>
<td>BLUE</td>
<td>BLUE</td>
<td>BLUE</td>
</tr>
</tbody>
</table>

E1) in order to estimate $\beta_i$, the estimator $\hat{\beta}_i$ needs $r$ variables, $\tilde{\beta}_i$ needs $q$ variables, while $\bar{\beta}_i$ needs $p$ variables (note that $p \leq r \leq q$).

E2) For any $M[i(p)]$ such that $M[i(p)] \subseteq M[t]$, $\hat{\beta}_i$ is a biased estimator of $\beta_i$, but $\tilde{\beta}_i$ and $\bar{\beta}_i$ are unbiased estimators of $\beta_i$.

E3) For any $M[i(p)]$ such that $M[i(p)] \supseteq M[t]$, $\tilde{\beta}_i$ which is equal to $\hat{\beta}_i$ is the Best Linear Unbiased Estimator (BLUE) of $\beta_i$, but $\bar{\beta}_i$ is not the BLUE. (Note that for any $M[i(p)]$ such that $M[i(p)] \supseteq M[t]$, $M[j(p)]$ of (4.6) coincides with $M[i(p)]$ and so $\tilde{\beta}_i = \hat{\beta}_i$ holds.)

E4) For $M[i(p)]$ such that $M[i(p)] = M[q]$, $\hat{\beta}_i$ and $\bar{\beta}_i$ are equal to $\hat{\beta}_i$ and $\tilde{\beta}_i$ is the BLUE.

E5) the estimator $\tilde{\beta}_i$ can not be determined in our situation because $M[t]$ is unknown. Note that from the properties E3) and E1), the estimator $\tilde{\beta}_i$ is better than $\bar{\beta}_i$ from the
point of not only estimation errors but also saving variables included in a model. From the above properties of the estimators \( \hat{\beta}_i, \tilde{\beta}_i \) and \( \tilde{\tilde{\beta}}_i \), the corresponding risk functions \( r(Y^*, \hat{\beta}_i), r(Y^*, \tilde{\beta}_i) \) and \( r(Y^*, \tilde{\tilde{\beta}}_i) \) have following properties (see Table 5):

<table>
<thead>
<tr>
<th>models</th>
<th>risk function</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( r(Y^*, \hat{\beta}_i) )</td>
<td>unestimable</td>
</tr>
<tr>
<td>( M[i(p)] \oplus M[t] )</td>
<td></td>
<td>(dependent on the situation)</td>
</tr>
<tr>
<td>( M[i(p)] \supseteq M[t] )</td>
<td>=</td>
<td>( \leq )</td>
</tr>
<tr>
<td>( M[i(p)] = M[q] )</td>
<td>=</td>
<td>=</td>
</tr>
</tbody>
</table>

R1) the first and second terms of \( r(Y^*, \hat{\beta}_i) \) and \( r(Y^*, \tilde{\beta}_i) \) are the same as those of \( r(Y^*, \hat{\beta}_i) \) because these terms are independent of the estimation method of \( \beta_i \),

R2) the third term of \( r(Y^*, \hat{\beta}_i) \) vanishes in \( r(Y^*, \tilde{\beta}_i) \) and \( r(Y^*, \tilde{\tilde{\beta}}_i) \) because the estimators \( \hat{\beta}_i \) and \( \hat{\beta}_i \) are unbiased estimators of \( \beta_i \) for any \( M[i(p)] \),

R3) from the property E3), the matrix \( V(\hat{\beta}_i) - V(\tilde{\beta}_i) \) is non-negative definite, and \( V(\tilde{\beta}_i) = V(\hat{\beta}_i) \) holds (where \( V(\tilde{\beta}_i) \) denotes the variance-covariance matrix of \( \tilde{\beta}_i \) and so on), therefore we have

\[
(6.1) \quad r(Y^*, \tilde{\beta}_i) = r(Y^*, \hat{\beta}_i) = r(Y^*, \tilde{\tilde{\beta}}_i)
\]

for any \( M[i(p)] \) such that \( M[i(p)] \supseteq M[t] \),

R4) in the models \( \{M[i(p)]; M[i(p)] \supseteq M[t]\} \), the relation of (6.1) does not hold (for example, in the model \( M[1(1)] \) of EX-1, the values of \( r(Y^*, \tilde{\beta}_i) \) and \( r(Y^*, \tilde{\beta}_i) \) are smaller than that of \( r(Y^*, \hat{\beta}_i) \)).

These properties are summarized in the Table-4 and Table-5. From the property E5), the criterion based on \( l(\tilde{\beta}_i) \) can not be used in our situation except in situation where \( M[t] \) is known. But from the properties E1), E3) and R3), the criterion based on \( l(\tilde{\beta}_i) \) is better than based on \( l(\tilde{\beta}_i) \). Therefore it is worthwhile to make the criterion based on \( l(\tilde{\beta}_i) \) applicable to practical situation.

7. Further Problem

Let us consider a procedure to make the criterion \( l(\tilde{\beta}_i) \) applicable to practical situation. The matrix \( W_{ij} \) of \( \tilde{\beta}_i \) can be obtained if \( M[j(r)] \) of (4.6) is determined. In the models \( \{M[k(s)]; M[k(s)] \supseteq M[i(p)]\} \), \( M[k*(s*)] \) coincides with \( M[j(r)] \) of (4.6) if the model \( M[k*(s*)] \) includes the true model \( M[t] \), i.e., if

\[
(7.1) \quad \{\beta_i' \Delta_0 \beta_i - \beta_i' \Delta_0 \beta_i\} = 0,
\]

and if \( s^* \) is the minimum in the models \( \{M[k*(s)]; M[k*(s)] \supseteq M[i(p)]\}, M[k*(s)] \supseteq M[t] \). The equality of (7.1) can be tested by the following unbiased estimator \( S_k \) of left-side of (7.1):
if we can find the distribution of $S_k$. Thus the model $M[j(r)]$ of (4.6) may be obtained. (In the section 5, the values of $r(Y^*, \hat{\beta}_i)$ were calculated under the assumption that $M[j(r)]$ is determined by this preliminary tests.) But there remains some problems to be considered, since the estimator $l(\tilde{\beta}_i)$ based on the above preliminary test may not be an unbiased estimator of $r(Y^*, \hat{\beta}_i)$ even if the distribution of $S_k$ is given. Then we will investigate the behavior of the estimator $l(\tilde{\beta}_i)$ based on the above preliminary test and the applicability of this criterion.

Appendix A.

Let us construct a matrix $W_{j|i}$ which satisfies (4.5). The singular value decomposition of $X_j$ and the basic value decomposition of $\Delta_H$ are expressed as follows:

\begin{align}
(a.1) & \quad X_j = H_j A_j^{1/2} G_j', \\
(a.2) & \quad \Delta_H = G_j^* A_j G_j^*, 
\end{align}

where $H_j$ is the $(n \times r)$ orthogonal matrix such that $H_j' H_j = I_r$, $A_j$ and $A_j^*$ are the $(r \times r)$ diagonal matrices in which diagonal elements are positive, and $G_j$, $G_j^*$ are $(r \times r)$ orthogonal matrices. Then (4.5) is expressed as

\begin{equation}
G_j A_j^{1/2} H_j' W_{j|i} H_j A_j^{1/2} G_j' = \Delta_H. 
\end{equation}

If the $(n \times n)$ matrix $W_{j|i}$ can be expressed using a $(r \times r)$ matrix $W_{j0}$ as

\begin{equation}
W_{j|i} = H_j W_{j0} H_j',
\end{equation}

$W_{j|i}$ is expressed as

\begin{align}
(a.5) & \quad W_{j|i} = H_j A_j^{1/2} G_j' A_j G_j A_j^{1/2} H_j', \\
& \quad = H_j A_j^{1/2} T_j A_j^{1/2} H_j', \\
& \quad = (A_j^{1/2} T_j A_j^{1/2}) A_j^{1/2} H_j', \\
& \quad = H_j A_j^{1/2} T_j A_j^{1/2} H_j', \\
& \quad = (A_j^{1/2} T_j A_j^{1/2}) A_j^{1/2} H_j',
\end{align}

where $T_j = G_j^* G_j$. From the equation

\begin{equation}
(X_j X_j)^{-1} X_j = G_j A_j^{1/2} H_j',
\end{equation}

the matrix $W_{j|i}$ of (a.5) is expressed as follows.

\begin{equation}
W_{j|i} = X_j (X_j X_j)^{-1} A_j (X_j X_j)^{-1} X_j
\end{equation}

And the matrix $W_{j|i}^{1/2}$ is expressed as follows using any $(n \times r)$ orthogonal matrix $H_j^*$ such that $H_j^* H_j^* = I_n$:

\begin{align}
(a.8) & \quad W_{j|i}^{1/2} = H_j^* A_j^{1/2} T_j A_j^{1/2} H_j', \\
& \quad = H_j^* A_j^{1/2} (X_j X_j)^{-1} X_j'.
\end{align}

Similarly, the matrix $W_\delta$ which satisfies (4.5)' is obtained as

\begin{align}
(a.9) & \quad W_\delta = X_\delta (X_\delta X_\delta)^{-1} A_\delta (X_\delta X_\delta)^{-1} X_\delta
\end{align}

And the matrix $W_\delta^{1/2}$ is expressed as follows using any $(n' \times q)$ orthogonal matrix $H_\delta^*$ such that $H_\delta^* H_\delta^* = I_n$:
Appendix B.

The equation $E_{\theta}(l(\hat{\theta}_i)) = r(Y^*, \hat{\theta}_i)$ is shown as follows. From (4.8) and $Z_i = W_{ji}^{1/2}X_i$ (where $W_{ji}^{1/2}$ is that of Appendix A), we have

$$Z_i\hat{\beta}_i = Z_i(X_i'W_{ji}X_i)^{-1}X_i'W_{ji}Y$$
$$= Z_i(Z_i'Z_i)^{-1}Z_i(W_{ji}^{1/2}Y)$$
$$= Q(Z_i)(W_{ji}^{1/2}Y).$$

Then

$$E_{\theta}(l(\hat{\theta}_i)) = E_{\theta}(l(\tilde{\theta}_i)) = r(Y^*, \tilde{\theta}_i).$$

From the assumption that $M[q] \supseteq M[\theta]$, we have

$$E_{\theta}(l(\hat{\theta}_i)) = E_{\theta}(l(\tilde{\theta}_i)).$$

Then

$$E_{\theta}(l(\hat{\theta}_i)) = E_{\theta}(l(\tilde{\theta}_i)).$$

Acknowledgment

The author is indebted to Dr. Y. Taga for drawing his attention to this problem. The author would like to acknowledge the continuing guidance and encouragement of Dr. M. Hoshi and Dr. M. Taguri. The author also wish to thank Dr. A. Hashimoto for his helpful advice in Section 4. Furthermore the author wishes to thank referees for their valuable comments.

References