Journal of the KSQC
Vol. 11, No. 1, 1983

The Integrated Bivar Criterion for Selecting Regressors

安 秉 珍*

ABSTRACT

A criterion is developed from the integrated sum of weighted squared bias and variance for use in selecting regressors. Some properties of this criterion are discussed and an example illustrating its use is included.

1. Introduction

This paper considers the problem of variable selection in a multiple regression. Choosing a subset involves considering the increased predictive variance caused by using too many regressors and increased bias caused by using too few.

There is considerable literature on the subject of the selection of the "best" subset of independent variables in a multiple regresson. Draper and Smith [1966], Mallows [1973], Helems [1974] and others propose criteria for variable selection. Hocking [1976] has reviewed criteria for variable selection. Many of these criteria are simple function of the mean squared error of estimation (or prediction), which trades off estimated variance against estimated squared bias on a one-to-one basis. Young [1982] propose the criterion that is based on a weighted sum of squared bias and variance in which the relative weights reflect the importance attached to these two quantities. The Park's criterion [1971] is concerned with precision not only at the design points but within some region of interest using weight function.

The proposed criterion in this paper is based on weighting integrated squared bias relative to integrated variance within some region of interest.

In section 2 we introduce the Integrated Bivar (Bias-variance) criterion and some properties of this criterion are discussed in section 3. In section 4 we propose plotting the Integrated Bivar criterion as a function of the weight in order to compare subsets and an example is given.

2. Formulation of the criterion

It is assumed that there are $n \geqq t+1$ observations on a t-vector of input variables, $x' = (x_1, ..., x_t)$, and a scalar response, y, such that

*서울대학교 계산통계학과 박사과정

the jth response, $j = 1,...,n$, is determined by

$$y_j = \beta_0 + \sum_{i=1}^{t} \beta_i x_{ij} + e_j . \qquad (2.1)$$

The residuals, e_j, are assumed identically and independently distributed, usually normal, with mean zero and unknown variance, σ^2. Note that implicit in these assumption is the assumption that the variables $x_1,...,x_t$ include all relevant variables although extraneous variables may be included.

The model (2.1) is frequently expressed in matrix notation as

$$Y = X\beta + e . \qquad (2.2)$$

Here Y is the n-vector of observed responses, X is the design matrix of dimension $n \times (t+1)$ as defined by (2.1), assumed to have rank $t+1$, and β is the $(t+1)$-vector of unknown regression coefficients.

Let the model (2.1) be written in matrix form as

$$Y = X_p \beta_p + X_r \beta_r + e , \qquad (2.3)$$

where the X matrix has been partitioned into X_p of dimension $n \times (p+1)$ and X_r of dimension $n \times r$. The β vector is partitioned conformably.

In the variable selection procedure, r is usually denoted the number of terms which are deleted from model (2.2) and $p = t + 1 - r$ is denoted the number of terms which are retained in the final equation.

The properties of residual mean squared error $\tilde{\sigma}^2$ and $\tilde{\beta}_p$ have been described by several authors.

We know that

$$E\left((n - p - 1)\,\tilde{\sigma}^2 \right) = E(SSE_p)$$

$$= (n - p - 1)\,\sigma^2 + \beta_r' X_r' \left[I - X_p (X_p' X_p)^{-1} X_p' \right] X_r \beta_r . \qquad (2.4)$$

If the subset model with X_r deleted is used, the estimated response is $\tilde{y}_p = x_p' \tilde{\beta}_p$ with mean

$$E(\tilde{y}_p) = x_p' \beta_p + x_p' A \beta_r \qquad (2.5)$$

where $A = (X_p' X_p)^{-1} X_p' X_r$

and $\mathrm{Var}(\tilde{y}_p) = x_p' (X_p' X_p)^{-1} x_p \sigma^2$. (2.6)

A risk function for judging the effectiveness of β_p as an estimator of β is given by

$$IH(w) = \frac{1}{\sigma^2} \int_R \left[\mathrm{Var}(\tilde{y}_p) + w\ bia^2s(\tilde{y}_p) \right] dW(x) \qquad (2.7)$$

where w is nonegative constant and R is region of interest.

In this $W(x)$ is measure defined over the Borel subsets of R satisfying: $\int_R dW(x) = 1$

$\int_R xx' dW(x) = M$, a finite matrix.

In practice, this means $W(x)$ can be treated as a probability distribution function with M as its matrix of second0order moments about orign.

From (2.5), (2.6) and (2.7),

$$IH = \frac{1}{\sigma^2} \int_R \left[x_p' (X_p' X_p)^{-1} x_p \sigma^2 + w \cdot \beta_r' \left[(x_p' A - x_r')' (x_p' A - x_r') \right] \times \beta_r \right] dW(x)$$

$$= tr\left[(X_p' X_p)^{-1} M_{pp} \right] + \frac{w}{\sigma^2} \beta_r' \left[A' M_{pp} A - 2A' M_{pr} \right.$$

$$+ M_{rr}] \beta_r \qquad (2.8)$$

where $M_{ij} = \int_R x_i x_j \, dW(x)$

and Tr denotes trace.

After replacement of parameters σ^2 and β_r by their estimators, the Integrated Bivar criterion is

$$Ic(w) = tr(X_p' X_p)^{-1} M_{pp} + \frac{w}{\hat{\sigma}^2} br'[A' M_{pp} A - 2A' M_{pr} + M_{rr}] b_r . \qquad (2.9)$$

Thus, the Integrated Bivar criterion is an estimator of $IH(w)$.

3. Properties of the proposed criterion

Park [1977] proposed by criterion which maximizes Q given by

$$\hat{Q} = \hat{\sigma}^2 \{ tr[(X'X)^{-1} M] - tr[(X_p' X_p)^{-1} M_{pp}] \} - b_r' [A' M_{pp} A - 2A' M_{pr} + M_{rr}] b_r \qquad (3.1)$$

where $M = \int_R xx' \, dW(x)$.

From (2.9) and (3.1)

$$Ic(w) = w \{ tr[(X'X)^{-1} M] - \frac{\hat{Q}}{\hat{\sigma}^2} \} + (1-w) tr[(X_p' X_p)^{-1} M_{pp}]$$

and $Ic(1) = tr[(X'X)^{-1} M] - \frac{\hat{Q}}{\hat{\sigma}^2}$

Thus, $Ic(1)$ rule is equivalent to max $\hat{Q}$ criterion.

If $W(x) = \frac{1}{n}$ at design points and $W(x) = 0$ elsewhere, then $M = (X'X)/n$.
From (2.4) and (2.8),

$$IH(w) = \frac{1}{n}[p + 1 + \frac{w}{\sigma^2} \beta_r' X_r' (I - X_p (X_p' X_p)^{-1} X_p') X_r \beta_r]$$

$$= \frac{1}{n}[p + 1 + \frac{w}{\sigma^2} (E(SSE_p) - (n-p-1)\sigma^2)] .$$

It follows that

$$Ic(w) = \frac{1}{n} \{ p + 1 + \frac{w}{\hat{\sigma}^2} SSE_p - w(n-p-1) \}$$

$$= \frac{1}{n} \{ w C_p + (1-w)(p+1) \}$$

where C_p is Mallow's C_p criterion, which is given by

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} + 2(p+1) - n .$$

Young [1982] proposed the criterion which is given by

$$C_I(w) = w C_p + (1-w)(p+1) .$$

Hence, If $W(x) = \frac{1}{n}$ at design points and $W(x) = 0$ elsewhere, $Ic(w)$ rule is equivalent to the rule proposed by Young.

Let's consider the special case of orthogonal regressors where $X'X = nI$.

$$Ic(w) = \frac{1}{n} tr[M_{pp}] + \frac{w}{\hat{\sigma}^2} b_r' M_{rr} b_r$$

$$= \frac{1}{n} tr\,[M_{pp}] + \frac{w}{n^2 \hat{\sigma}^2} \cdot y' X_r$$

$$M_{rr}\ X_r' y \qquad (3.2)$$

(case 1) If $W(x) = \frac{1}{n}$ at the design points and $W(x) = 0$ elsewhere, then

$$M = (X'X)/n = I.$$

We know that

$$\hat{\sigma}^2 = Y'(I - \frac{1}{n}(X_p X_r'))$$

$$-\frac{1}{n}(X_r X_r'))\, Y/(n - p - 1)$$

and $$SSE_p = Y'(I - \frac{1}{n} X_p X_p')\, Y.$$

Equation (3.2) becomes

$$Ic(w) = \frac{p+1}{n} + w \cdot Y' X_r X_r' Y / n^2 \hat{\sigma}^2$$

$$= \frac{1}{n} \{ p + 1 + w (\frac{SSE_p}{\hat{\sigma}^2} - n$$

$$+ p + 1) \}.$$

Let $Ic(w|p)$ denote the $Ic(w)$ for given p.

$$Ic(w|p+1) - Ic(w|p)$$

$$= \frac{1}{n} \{ \frac{w}{\hat{\sigma}^2} (SSE_{p+1} - SSE_p) + (1 + w) \}$$

$$= \frac{1}{n} \{ -w t_\nu^2 + (1 + w) \}$$

where t_ν^2 is the t statistic with $\nu = n - t - 1$ degrees of freedom. Thus additional regressor is included if and only it $t_\nu^2 > (1 + 1/w)$. In case of $M = I$, $Ic(w)$ rule is equivalent to sequential t tests on regression coefficient with critical t values at $\sqrt{1 + 1/w}$

(case 2) If $W(x)$ be uniform weight function over the region of interest R,

$$M = \int_R x x' dW(x) = \frac{\int_R x x' dx}{\int_R dx}$$

$$= \begin{bmatrix} 1 & \vdots & \underset{\sim}{0}' \\ \cdots & \cdots & \cdots \\ \underset{\sim}{0} & \vdots & \frac{1}{3} I \end{bmatrix}$$

where $R = \{ (x_1, \cdots\cdots, x_k)$

$$; |x_i| \leq 1 \quad i = 1, 2, \cdots\cdots, k \}$$

Equation (3.2) becomes

$$Ic(w) = \frac{1}{n} (1 + \frac{p}{3})$$

$$+ w \cdot Y' X_r X_r Y / 3n^2 \hat{\sigma}^2$$

$$= \frac{1}{3n} \{ 3 + p + w \cdot (\frac{SSE_p}{\sigma^2} - n$$

$$+ p + 1) \}.$$

Thus, we get the same result as case 1.

4. The use of Ic(w) plots.

We can use $Ic(w)$ plots like the $C_I(w)$ plots suggested by Young [1982]. For fixed subset of variables, $Ic(w)$ is the first order linear function of w. A subset is considered optimal at a point w if its line is lower than all other eines at that point.

The $Ic(w)$ plot is used to determine the sub-

set that is optimal for a wide interval of moderate values of w. If such an interval ranges from zero to well beyond one, then the subset is "optimal resistent" to changes in relative concern both for predicting and for esfimating mean response.

When there is no such subset, then the subset that is optimal for values of w close to zero is best for prediction. While the optimal subset for large values of w is best for estimating mean response. The subset with optimal range between these and including $w = 1$ is a good compromise subset that can be used for both purposes.

The $Ic(w)$ plot thus offers a complete picture of alternatives and conditions under which each alternative is optimal. The final selection would then depend on the conditions that are most desiable for a particular problem.

Example. The gas mileage data that discussed by Hocking [1976] to predict gasoline mileage for 1973-1974 automobiles, road tests were performed by "Mortor Trend" magzine in which gasoline mileage and ten physical characteristics of various type of automobiles were recorded.

TABLE 1.

Description of variables for the gas mileage data (32 observations)

Number	Description
x_1	Engin Shape
x_2	Number of Cylinders
x_3	Transmission Type
x_4	Number of Transmission Speeds
x_5	Engine Size
x_6	Horse power
x_7	Number of Carburetor barrels
x_8	Final Drive Ratio
x_9	Weight (pounds)
x_{10}	Quarter Mile Time (Seconds)
y	Gasoline Mileage

For simplicity, the data are standardized as following:

$$\sum_{j=1}^{32} x_{ij} = 0, \quad \sum_{j=1}^{32} x_{ij}^2 = 1$$

$$i = 1, 2, \cdots\cdots, 10$$

$$\sum_{j=1}^{32} y_j = 0, \quad \sum_{j=1}^{32} y_j^2 = 1$$

We will take the weight function, $W(x)$, to be uniform over the region of interest $R = \{(x_1, x_2, \ldots, x_{10}); |x_i| \leq 1, i = 1,2,\ldots,10\}$. A summary of the relevant quantities is given in TABLE 2.

TABLE 2.

Summary of the relevant quantities for gas-mileage data

p	$MSE_p \times 10^3$	C_p	$\hat{Q} \times 10^3$	$Ic(w) \times 10$
2	5.85	1.22	125	9.98w + 1.18
3	5.37	1.10	133	8.09w + 2.13
4	5.26	0.79	139	4.35w + 4.88
5	5.24	1.85	134	0.68w + 9.34
6	5.33	3.37	130	0.07w + 10.65
7	5.50	5.15	115	-1.10w + 13.32
8	5.71	7.05	99.1	-0.04w + 15.92
9	5.96	9.10	60.8	0.13w + 22.13

The $Ic(w)$ function for different p are plotted in Figure 1. Figure 1 shows that the optimum number of terms retained is different for different values of w.

Note that i) For values of w close to zero $p = 3$ is disirable selection, which

is the same result as C_p criterion.

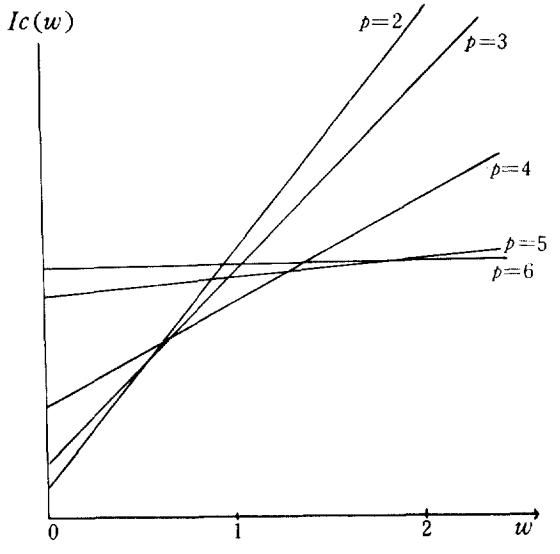
ii) For large values of w $p = 5$ is disirable selection, which is the same result as MSE_p criterion.

iii) For values of w including one $p = 4$ is disirable selection, which is the same result as $\hat{Q}$ criterion.

Thus, $p = 3$ is preferable for prediction and $p = 5$ for estimation and $p = 4$ is desiable for both purposes.

Figure 1.

$Ic(w)$ plots for Gasmileage data.



5. Summary

We have proposed a selection criterion which gives a picture of alternative regressor subsets and condition under which each alternative is optimal for given $W(x)$. The criterion is based on weighting integrated squared bias relative to integrated variance over the region of interest R. If $W(x) = \frac{1}{n}$ at design points and $W(x) = 0$ elsewhere, $Ic(w)$ rule is equivalent to the rule proposed by Young [1982] and $Ic(1)$ rule is equivalent to the rule proposed by Park [1977].

The use of this criterion was proposed through $Ic(w)$ plot and an example of such an use was given.

6. Acknowledgement

The author is grateful to Professor Sung H. Park for suggesting this problem and invaluable guidence in the development of this paper.

References

Draper, N.R. and Smith, H., (1966). "Applied Regression Analysis", Wiley, New York.

Mallows, C., (1973). "Some comments on C_p", Technometrics, Vol. 15, pp. 661-675.

Helms, R., (1974). "The average estimated variance criterion for the selection of variables problem in general linear models", Technometrics, Vol. 16, pp. 261-274.

Hocking, R.R., (1976). "The analysis and selection of variables in linear regression", Biometrics, Vol. 32, pp. 1-49.

Park, S.H., (1977). "Selection of polynomial terms for response surface experiments", Biometrics, Vol. 33, pp. 225-229.

_____, (1978). "Selecting contrasts among parameters in scheffe's mixture models: Screening components and reduction", Technometrics, Vol. 20, pp. 273-279.

Young, A.S., (1982). "The Bivar Criterion for selecting Regressors", Technometrics, Vol. 24, pp. 181-189.