Faculty Working Papers

INFORMATION CRITERIA FOR THE CHOICE OF REGRESSION MODELS

Takamitsu Sawa

#368

College of Commerce and Business Administration
University of Illinois at Urbana-Champaign

FACULTY WORKING PAPERS

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

January 4, 1977

INFORMATION CRITERIA FOR THE CHOICE OF REGRESSION MODELS

Takamitsu Sawa

#368

Information Criteria for the
Choice of Regression Models

Takamitsu Sawa

Associate Professor, Economic Department

January 17, 1977

1. Introduction

In most statistical analyses it is taken for granted that the family of the probability distribution functions, say $F(y|\theta)$, may be correctly specified on *a priori* grounds. Uncertainty exists, therefore, only with reference to the values of parameters θ involved in the specified family of probability distribution functions (p.d.f.). In practice, however, we are seldom in such an ideal situation; that is, we are more or less uncertain about the family to which the *true* p.d.f. might belong. It may be very likely that the true distribution is in fact too complicated to be represented by a simple mathematical function such as is given in ordinary textbooks.

In practice we approximate the true distribution by one of the alternative p.d.f.'s listed in the textbooks. Needless to say, we try to choose the most adequate p.d.f. with due thought to *a priori* considerations. The p.d.f. specified by a convenient mathematical function is usually termed the *model*. For further analysis the model is identified at least tentatively with the true distribution. To put it differently, in the process of conventional statistical analysis a sharp distinction is seldom drawn between the model and the true distribution.

To avoid the arbitrariness that inevitably occurs in the process of model building, nonparametric statistical methods have been extensively developed in the past decade. It seems to me, however, that these methods have not been used very successfully in practical data analysis.

In fact, most statistical inferences are based on some specific parametric model, often on the model of normal distribution.

In recent years, however, more and more emphasis has been laid on the problem of model identification; that is, how to identify the model when it cannot be completely specified from a priori grounds. The main purpose of the present paper is to propose and analyze a statistical criterion for model identification in regression analysis. Our basic attitude toward the problem is to recognize the fact that a certain amount of discrepancy inevitably exists between the true distribution and the model. The best we can do in trying to cope with this sort of situation is to identify the most adequate model among a given set of alternatives. The adequacy of a model needs to be quantified by introducing a suitable measure of the distance of the model from the unknown true distribution.

It is expected intuitively that the more complicated model will provide the better approximation to reality. But, on the contrary, the less complicated model should be preferred if we wish to pursue accuracy of estimation. To illustrate this point, let us consider the situation where two alternative density functions $f_1(\cdot|\theta)$ and $f_2(\cdot|\zeta)$, are given as possible models of the density $g(\cdot)$ of the true distribution, where θ and ζ are vectors of unknown parameters. Even if $f_1(\cdot|\theta)$ is the better approximation to the true density $g(\cdot)$ in the sense that $\inf_\theta \| f_1(\cdot|\theta) - g(\cdot)\| < \inf_\zeta \| f_2(\cdot|\zeta) - g(\cdot)\|$ where $\|\cdot\|$ is a suitably defined distance, it is quite likely that $E_{\hat\theta} \| f_1(\cdot|\hat\theta) - g(\cdot)\| > E_{\hat\zeta} \| f_2(\cdot|\hat\zeta) - g(\cdot)\|$ if dim $\theta >$ dim ζ where $\hat\theta$ and $\hat\zeta$ are estimates for θ and ζ respectively.

The above consideration leads us naturally to the so-called principle of parsimony. That is, more parsimonious use of parameters should be pursued so as to raise the accuracy of the estimates of the parameters. In general, closeness to the true distribution is incompatible with parsimony of parameters. These two criteria form a trade-off. That is, if one pursues one of the criteria, the other must be necessarily sacrificed. The multiple correlation coefficient adjusted for the degrees of freedom may be the most commonly used statistic that incorporates the two incompatible criteria into a single statistic.

Akaike [1] has proposed a more general as well as more widely applicable statistic that ingeniously incorporates the two criteria. Since it is based on the Kullback-Leibler Information Criterion, Akaike's statistic is called the Akaike Information Criterion and is abbreviated as AIC. Indeed, the procedure developed here is also based on the Kullback-Leibler Information Criterion, but the criterion for the choice of a regression model implied by our procedure is considerably different from that implied by AIC. The disagreement stems from a difference between Akaike's and our views on the true distribution.

In Section 2 we briefly review the Kullback-Leibler Information Criterion and the Akaike Information Criterion. In Section 3 we develop a criterion for the choice of a regression model and compare it with a criterion implied by the Akaike Criterion. In Section 4 the Bayesian approach to the problem is considered and a different criterion is derived from Bayesian point of view. The bias of the three criteria is discussed in Section 5.

2. Information Criterion

Suppose that we are concerned with the probabilistic structure of a vector random variable $Y' = (Y_1, Y_2, \cdots, Y_n)$. Let $G(y)$ be the true joint distribution of Y. On the basis of a priori knowledge we postulate a model $F(y|\theta)$ to approximate the unknown true distribution $G(y)$, where θ is a vector of unknown parameters.

The adequacy of a postulated model may be measured by the Kullback-Leibler's Information Criterion (KLIC).

$$(2.1) \qquad I(G:F(\cdot|\theta)) = E_G[\log \frac{g(Y)}{f(Y|\theta)}]$$

$$= \int \log \frac{g(y)}{f(y|\theta)} \, dG(y)$$

where g and f are density (or probability) functions of, respectively, G and F; $E_G(\cdot)$ stands for expectation with respect to the true distribution G; the integration is over the entire range of Y. It can be easily shown that the KLIC is nonnegative

$$(2.2) \qquad I(G:F(\cdot|\theta)) \geq 0$$

with equality only when $F(y|\theta) = G(y)$ almost everywhere in the possible range of Y; namely, only when the model is correct. (See, for instance, Rao [5] pp. 58-59.) Incidentally, the negative value of the KLIC is termed the entropy of a probability distribution $G(y)$ with respect to $F(y|\theta)$. Noting the inequality (2.2) as well as an obvious equality

$$(2.3) \qquad I(G:F(\cdot|\theta)) = \int \log g(y)dG(y) - \int \log f(y|\theta)dG(y),$$

we are led to propose the following rule for a comparison of alternative models or estimates.

Rule 2.1: (i) A model $F_1(\cdot|\theta)$ is regarded as a better approximation to the true distribution $G(\cdot)$, i.e., a better model than an alternative model $F_2(\cdot|\zeta)$ if and only if

$$(2.4) \qquad \inf_{\theta} I(G:F_1(\cdot|\theta)) < \inf_{\zeta} I(G:F_2(\cdot|\zeta))$$

or equivalently

$$(2.5) \qquad \sup_{\theta} E_G [\log f_1(Y|\theta)] > \sup_{\zeta} E_G[\log f_2(Y|\zeta)].$$

(ii) Given a model $F(\cdot|\theta)$, an estimate $\hat{\theta}_1$ is regarded as a better estimate than $\hat{\theta}_2$, if and only if

$$(2.6) \qquad E_{\hat{\theta}_1}\{E_G[\log f(Y|\hat{\theta}_1)|\hat{\theta}_1]\} > E_{\hat{\theta}_2}\{E_G[\log f(Y|\hat{\theta}_2)|\hat{\theta}_2]\}$$

where $E_{\hat{\theta}_1}$ and $E_{\hat{\theta}_2}$ stand for expectations with respect to the sampling distributions of $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively. (Note that when we first take an expectation with respect to G the estimate $\hat{\theta}_1$ or $\hat{\theta}_2$ should be treated as if it were a constant.)

It was pointed out by Akaike [1] that if the $Y_j^!$ s are independent and identically distributed the maximum likelihood estimate may be regarded as an estimate that minimizes the estimated KLIC, or equivalently maximizes the estimated entropy, because the log likelihood function divided by the sample size n

$$(2.7) \qquad \frac{1}{n} \sum_{j=1}^{n} \log f(y_j|\theta)$$

may be regarded as a reasonable estimate for $E_G[\log f(Y|\theta)]$ whatever $G(y)$ is.

Apparently, the above rule for a comparison of models is not directly applicable in practice, because the criteria are totally dependent on the unknown true probability distribution. To establish a practical usable criterion for model identification on the basis of the KLIC, we need to replace unknowns in (2.5) by their reasonable estimates. In fact, the Akaike Information Criterion (AIC) has been derived as an approximately unbiased estimate for the KLIC, neglecting its irrelevant constant terms and based implicitly on a fairly strong assumption.

For the sake of convenience in developing our argument we give the following definition:

Definition: Given a model $F(\cdot|\theta)$, a parameter value θ_0 such that

(2.8) $$I(G:\ F(\cdot|\theta_0)) \leq I(G:\ F(\cdot|\theta))$$

for any possible θ in the admissible parameter space is called a pseudo-true parameter value.

If the true distribution $G(y)$ and a model $F(y|\theta)$ satisfy due regularity conditions, the pseudo-true parameter θ_0 must satisfy

(2.9) $$E_G[\frac{\partial}{\partial\theta} \log f(Y|\theta)]_{\theta=\theta_0} = 0.$$

The model $F(y|\theta_0)$ may be regarded as the most adequate relatively within the family of models $F(y|\theta)$ in the sense that the KLIC for $F(y|\theta)$ is minimized by $F(y|\theta_0)$.

Assuming that $G(y) = F(y|\theta_0)$ almost everywhere, Akaike [1] derives his criterion

$$\text{(2.10)} \qquad \mathrm{AIC}\,(F(\cdot|\theta)) = -2 \log f(y|\hat{\theta}) + 2k$$

as an almost unbiased estimate for $-2\, E_G\, [\log f(Y|\theta)]$, where $\hat{\theta}$ is the maximum likelihood estimate for θ based on observations y and k is the number of the unknown parameters, i.e., the dimension of θ. The procedure of choosing a model that minimizes the AIC is called the Minimum AIC (MAIC) procedure. The first term of the AIC measures the goodness of fit of the model to a given set of data, because $f(y|\hat{\theta})$ is the maximized likelihood function. The second term is interpreted as representing a penalty that should be paid for increasing the number of parameters. The increase in the number of parameters almost necessarily improves the fit but only at the cost of sacrificing accuracy of estimation. In this sense the AIC may be regarded as an explicit formulation of the so-called principle of parsimony in model building.

Indeed, the assumption that

$$\text{(2.11)} \qquad F(y|\theta_0) = G(y)$$

simplifies the derivation substantially, but there is no denying that this simplifying assumption lessens the plausiblity of the AIC to some extent. In the next section, confining ourselves to a linear regression, we derive another criterion without assuming (2.11) and compare it with the AIC to see what difference might arise depending on whether or not we assume (2.11).

3. Identification of a Regression Model

We are interested in investigating a joint distribution of a vector random variable $Y' = (Y_1, Y_2, \cdots, Y_n)$. Each of Y_i's may be an observation on a certain characteristic of a randomly chosen individual; or Y_i's may constitute a sequence of observed time series. The distribution function $G(y)$ is unknown, but each Y_i is assumed to possess finite variance. We denote the mean vector and the variance-covariance matrix, respectively, by μ and Ω, where μ is a vector of n components and Ω is a $n \times n$ positive definite matrix. Unless we place more a priori restrictions on the elements of μ and Ω, we can make no inference at all about the joint distribution of Y.

What we usually do is to assume that μ belongs to a linear subspace of lower dimension than n and Y_i's are mutually uncorrelated. Then we have a familiar linear regression model

$$(3.1) \qquad E(Y) = X\beta, \; V(Y) = \sigma^2 I_n$$

where X is a $n \times k$ matrix of known constants, the k columns of which constitute a basis of the subspace to which μ is assumed to belong; β is a vector of k unknown parameters; σ^2 is an unknown positive constant; I_n is an identity matrix of order n. In most practical situations the columns of X are vectors of observations on certain characteristics considered to be associated with Y. Then the model implies that the i-th mean μ_i is represented as a linear function of

k explanatory variables, i.e., $\mu_i = \sum_{j=1}^{k} \beta_j x_{ij}$ where x_{ij} is the (i, j)-th element of x. By assuming a regression model we can reduce the number of unknown parameters from $n + n(n + 1)/2$ to $k + 1$.

In addition to (3.1) we often assume the normal distribution for Y, and postulate a model

$$Y \sim N(X\beta, \sigma^2 I_n), \tag{3.2}$$

or

$$Y = X\beta + u, \qquad u \sim N(0, \sigma^2 I_n),$$

which is termed a linear normal regression model.

Lemma 3.1: The pseudo true values for parameters $\theta' = (\beta', \sigma^2)$ are

$$\beta_0 = (X'X)^{-1}X'\mu \tag{3.3}$$

$$\sigma_0^2 = \frac{1}{n} \mu'(I - X(X'X)^{-1}X')\mu + \frac{1}{n} \operatorname{tr} \Omega. \tag{3.4}$$

The above results are easily obtained by solving the equations

$$E[\frac{\partial}{\partial\beta} \log f(Y|\theta)] = 0 \tag{3.5}$$

$$E[\frac{\partial}{\partial\sigma^2} \log f(Y|\theta)] = 0 \tag{3.6}$$

where $f(y|\theta)$ is the density function of $N(X\beta, \sigma^2 I)$ and the expectation is with respect to the true distribution. Geometrically speaking, $X\beta_0$ is a projection of the unknown mean vector μ into the space spanned by the k columns of X, while $n\sigma_0^2$ is the sum of the variances of the Y_j's plus the squared length of the perpendicular from μ to the space. The error of approximating μ by $X\beta$ is absorbed into the error variance.

The maximum likelihood (ML) estimates

$$(3.7) \qquad \hat{\beta} = (X'X)^{-1}X'y, \qquad \hat{\sigma}^2 = \frac{1}{n} y'[I - X(X'X)^{-1}X']y$$

for β and σ^2 in the normal regression model (3.2) have the following property.

Lemma 3.2

$$(3.8) \qquad E(\hat{\beta}) = \beta_0,$$

$$(3.9) \qquad \text{plim}\ (\hat{\sigma}^2 - \sigma_0^2) = 0, \text{ if } \Omega = \omega^2 I_n$$

This lemma implies that with an incorrect model the objects of our estimation are pseudo true parameter values. To put it differently, what we ordinarily call the true parameter values are the parameter values that minimize the distance between the true unknown distribution and the postulated parametric model, where the distance is measured by the KLIC. Moreover, it should be noted that if Y_i's are uncorrelated, i.e., $\Omega = \omega^2 I_n$, then $\hat{\beta}$ and $\hat{\sigma}^2$ are uncorrelated.

Along the lines of the previous section, one can measure the loss incurred by modelling $G(y)$ by $F(y|\tilde{\theta})$ with some estimate $\tilde{\theta}$ in place of unknown θ_0 by the quantity

$$(3.10) \qquad W(F(\cdot|\tilde{\theta})) = -\frac{2}{n} E_G [\log f(Y|\tilde{\theta})|\tilde{\theta}],$$

where $f(y|\theta)$ is the density function of the pseudo-true model $N(X\beta_0, \sigma_0^2 I)$, i.e., the likelihood function of the model. It should be noted that the expectation on the right-hand side of (3.10) refers only to the argument Y of the density function.

Lemma 3.3:

The loss incurred by modelling the distribution of Y by $F(y|\tilde{\theta})$ with an estimated value $\tilde{\theta}$ substituted for θ is evaluated as

$$W(F(\cdot|\hat{\theta})) = \log(2\pi) + \log(\tilde{\sigma}^2) + (\frac{\sigma_0^2}{\tilde{\sigma}^2}) + \frac{1}{n\tilde{\sigma}^2} \| X(\tilde{\beta} - \beta) \|^2, \tag{3.11}$$

where $\| \cdot \|$ is the Euclidean norm.

The proof is given in the Appendix.

In this section we adhere to the sampling theory approach, and hence we base our decision about model selection on the risk function derived by integrating the loss function with respect to the sampling distribution of the estimate $\tilde{\theta}$. Since the ML estimate $\hat{\theta}$ possesses the nice property in Lemma 3.2, even when a postulated model is incorrect, we define the risk of postulating a model $F(y|\theta)$ by an integral of the loss function of $F(y|\hat{\theta})$ with respect to the sampling distribution of the ML estimate $\hat{\theta}$.

Theorem 3.1: Suppose that $\Omega = \omega^2 I_n$ and each Y_j is symmetrically distributed with the same kurtosis as a normal distribution. Then the risk of a model $F(\cdot|\theta)$, i.e., the expected value of $W(F(\cdot|\hat{\theta}))$, is evaluated to order $O(n^{-1})$ as

$$R(F(\cdot|\theta)) = \log(2\pi) + \log(\sigma_0^2) + 1 + \frac{k+2}{n}(\frac{\omega^2}{\sigma_0^2}) - \frac{1}{n}(\frac{\omega^2}{\sigma_0^2})^2 + O(n^{-2}).$$

The proof is given in the Appendix. It should be noted that σ_0^2 increases with the addition of explanatory variables, i.e., the increase of k.

To develop a practical and useful criterion for model identification, the risk function involving unknown parameters needs to be somehow estimated from a given set of observations.

Theorem 3.2: Suppose that an asymptotically unbiased estimate, say $\hat{\omega}^2$, for ω^2 is obtained from some source available a priori and it is statistically independent of $\hat{\sigma}^2$. Then

$$\text{(3.13)} \qquad \text{BIC}\ (F(\cdot|\theta)) = -2 \log f\ (y;\hat{\theta}) + 2(k+2)\left(\frac{\hat{\omega}^2}{\hat{\sigma}^2}\right) - 2\left(\frac{\hat{\omega}^2}{\hat{\sigma}^2}\right)^2$$

is an asymptotically unbiased estimate of $nR(F(\cdot|\theta))$.

The proof is given in the Appendix. If we equate $\hat{\omega}^2$ to $\hat{\sigma}^2$, the BIC is identical with the AIC. As was pointed out in the preceding section, the AIC is based on the assumption that the true distribution belongs to the family of distributions specified by a postulated model; namely, σ_0^2 is equated to ω^2 in the process of deriving the AIC.

The variance ratio $\hat{\omega}^2/\hat{\sigma}^2$ increases with successive addition of explanatory variables, and possibly it approaches one. Its reciprocal $\hat{\sigma}^2/\hat{\omega}^2$ (≥ 1) may be interpreted as a discounting factor for the penalty that has to be paid for increasing the number of parameters. Therefore, when we compare two regression models, one with less explanatory variables and poorer fit, the other with more explanatory variables and better fit, the BIC is more favorable to the more parsimonious model than the AIC. The following numerical evaluations show that the difference between the two criteria is far from negligible.

Let us develop a decision rule to choose one from two alternative regression models

$$F_1: \quad Y \sim N(X_1\beta_1, \sigma_1^2 I_n)$$

$$F_2: \quad Y \sim N(X_1\beta_1 + X_2\beta_2, \sigma_2^2 I_n)$$

where X_1 and X_2 are respectively n x p and n x q matrices of known constants, β_1 and β_2 are respectively p x 1 and q x 1 vectors of unknown parameters, and σ_1^2 and σ_2^2 are positive unknowns. The true distribution is assumed to be $N(\mu, \omega^2 I_n)$. In practice, we cannot expect to obtain an estimate for ω^2 from some independent source. Therefore, assuming that the more complicated model F_2 is nearly true, we substitute the ML estimate $\hat{\sigma}_2^2$ of σ_2^2 for $\hat{\omega}^2$ in (3.13). Our decision rule is described as follows: we choose F_1 if BIC (F_1) < BIC (F_2) and vice versa.

It is straightforward to show that the decision rule based on the BIC is equivalent to a decision based on the magnitude of the F-statistic that is customarily used to test the null-hypothesis $\beta_2 = 0$. That is, we decide to choose F_1 if an observed value of the F-statistic falls below a critical point determined by the inequality, BIC (F_1) < BIC (F_2) and choose F_2 otherwise. The critical point varies depending on n, p, and q.

Confining ourselves to the case when q = 1, we tabulate the critical points implied by the minimum BIC principle, say MBIC critical points, in Table 3.1. Since the t-statistic is more familiar to us than the F-statistic in the case of q = 1, these critical values refer to the t-statistic. We decide to choose F_1 if the observed value of the t-statistic, the ML estimate of β_2 divided by its estimated standard deviation, falls below a critical point read from the table, and vice versa.

To examine how much the MBIC procedure differs from the MAIC procedure, the critical points implied by the AIC are also tabulated in Table 3.2. Both of these approach, although very slowly,

$\sqrt{2}$ asymptotically. We note a remarkable difference, namely that the MAIC critical point approaches $\sqrt{2}$ from below whereas the MBIC approaches from above. Moreover, as the number of variables already included increases, i.e., as p becomes larger, the MBIC procedure increasingly discriminates against the inclusion of additional variables, whereas the converse is true for MAIC.

To see a connexion between our procedure and the preliminary t-test, for some chosen cases, we tabulate the level of significance, i.e., the probability that $|t|$ exceeds the critical point when F_1 is true. Roughly speaking, for moderate values of p, the significance level for the MAIC procedure varies over the wide range from 30% to 16% as the number of degrees of freedom increases; on the other hand, for the MBIC procedure, it varies over a relatively narrow range from 10% to 16%. Both procedures share a common property in their more generous attitude toward inclusion of additional variables than the traditional preliminary test with the significance level 5% or 10%. It should be noted, however, that these two asymptotically equivalent procedures will very often lead us to different decisions for small samples.

Based on the minimax regret principle with the squared error of prediction as a loss function, Sawa and Hiromatsu [6] calculated the optimal significance point for the preliminary t-test. Their minimax regret significance points are quite insensitive to the change in the number of degrees of freedom. That is, it remains constant at 1.37 to two decimal places, unless the number of degrees of freedom is extremely small, say less than 10. Indeed it is difficult to establish a clear-cut connection between the two basically different approaches, but it would

be worth noting that if a loss function is specified in terms of the prediction error, the more prodigal model is likely to be preferred.

We often encounter a situation where we have to choose one of two unnested alternatives:

$$Y \sim N\ (X_1\beta_1,\ \sigma_1^2 I_n) \text{ and } Y \sim N\ (X_2\beta_2,\ \sigma_2^2 I_n),$$

where the true distribution of Y is N $(\mu,\ \omega^2 I_n)$. In this kind of situation the unknown true variance ω^2 may be reasonably estimated from a regression of y on all the explanatory variables $X_1 \cup X_2$. Another reasonable estimate of ω^2 may be the smallest value of "unbiased" estimates of variances for all possible regressions of y on a subset of $X_1 \cup X_2$.

The difficulty in estimating ω^2 does admittedly place a serious limitation to the practical usefulness of the MBIC procedure. However, it should be noted that the same difficulty is shared by Mallow's [4] procedure which is based on what he calls C_p statistic. Incidentally, Mallow's procedure gives a decision rule essentially similar to the AIC. It is also worth noting that according to Akaike's procedure ω^2 is estimated by $\hat{\sigma}_1^2$ when we evaluate the AIC for the model F_1 and by $\hat{\sigma}_2^2$ when we evaluate the AIC for the model F_2. This means that, given a class of nested alternative models, the AIC for each model is evaluated assuming it is true. On the other hand, the BIC for each model is evaluated assuming that the most complex model within the class would be true.

Table 3.1

MBIC Critical Points for the Preliminary t-Test

p n	1	2	3	4	5	10
10	1.525	1.646	1.816	2.036	2.264	-
12	1.500	1.591	1.715	1.882	2.092	-
14	1.484	1.557	1.652	1.778	1.943	2.678
16	1.473	1.533	1.610	1.709	1.836	2.758
18	1.465	1.516	1.580	1.660	1.761	2.665
20	1.458	1.504	1.558	1.625	1.707	2.494
25	1.448	1.482	1.522	1.568	1.624	2.192
30	1.442	1.469	1.500	1.536	1.576	1.912
50	1.430	1.445	1.462	1.480	1.449	1.625
100	1.442	1.429	1.437	1.445	1.453	1.499
200	1.418	1.421	1.425	1.429	1.433	1.453
500	1.416	1.417	1.419	1.420	1.421	1.429
1000	1.415	1.416	1.416	1.417	1.418	1.421

n is the sample size and p_1 is the number of the explanatory variables already included in the model. The decision rule is described as follows: if the t-value for an optimal variable exceeds the MBIC critical point, we decide to augment the model by the optimal variable, and *vice versa*. Note that the MBIC critical point approaches slowly to $\sqrt{2}$ as n tends to infinity for every p.

Table 3.2

MAIC Critical Points for the Preliminary t-Test

p n	1	2	3	4	5	10
10	1.331	1.245	1.153	1.052	.941	-
12	1.346	1.278	1.205	1.127	1.043	.426
14	1.357	1.300	1.239	1.176	1.108	.679
16	1.365	1.316	1.264	1.210	1.154	.816
18	1.371	1.328	1.283	1.236	1.188	.907
20	1.376	1.337	1.297	1.256	1.213	.973
25	1.384	1.354	1.323	1.291	1.258	1.000
30	1.389	1.364	1.339	1.313	1.286	1.144
50	1.400	1.385	1.370	1.355	1.340	1.262
100	1.407	1.400	1.393	1.385	1.378	1.341
200	1.411	1.407	1.404	1.400	1.396	1.378
500	1.413	1.411	1.410	1.409	1.407	1.400
1000	1.414	1.413	1.412	1.411	1.411	1.407

n is the sample size and p is the number of the explanatory variables already included in the model. The decision rule is described as follows: if the t-value for an optional variable exceeds the MAIC critical point, we decide to augment the model by the optional variable, and *vice versa*. Note that the MAIC critical point approaches slowly to $\sqrt{2}$ as n tends to infinity for every p.

Table 3.3 Significance Levels Implied by the BIC Procedure

n ^{p}	1	2	3	4	5
10	.1658	.1438	.1193	.0974	.0863
12	.1645	.1461	.1247	.1019	.0814
14	.1636	.1478	.1295	.1091	.0879
16	.1629	.1492	.1334	.1155	.0962
18	.1625	.1503	.1364	.1208	.1037
20	.1621	.1509	.1388	.1250	.1099
25	.1611	.1525	.1430	.1326	.1209
30	.1604	.1534	.1457	.1371	.1281
50	.1592	.1551	.1505	.1458	.1544
∞	.1574	.1574	.1574	.1574	.1574

Table 3.4 Significance Levels Implied by the AIC Procedure

n ^{p}	1	2	3	4	5
10	.2199	.2532	.2928	.3410	.4000
12	.2080	.2332	.2626	.2969	.3371
14	.1998	.2202	.2436	.2698	.3001
16	.1938	.2109	.2302	.2516	.2753
18	.1893	.2040	.2203	.2383	.2578
20	.1857	.1988	.2130	.2283	.2452
25	.1796	.1895	.2001	.2114	.2236
30	.1758	.1838	.1922	.2011	.2107
50	.1679	.1726	.1773	.1822	.1871
∞	.1574	.1574	.1574	.1574	.1574

4. Bayesian Decision Rule

In this section we look at the problem another way, from the Bayesian point of view. Given a model $F(\cdot|\theta)$ coupled with a prior distribution $P(\theta)$ we define the Bayes risk, say $B(\tilde{\theta}|F)$, for an estimate $\tilde{\theta}$ by the expectation of the loss function (3.10) with respect to the posterior distribution, that is,

$$(4.1) \qquad B(\tilde{\theta}|F) = \int W(F(\cdot|\tilde{\theta}))\, dP(\theta|y)$$

where $P(\theta|y)$ is the posterior distribution for θ given an observation y. If there exists an estimate $\tilde{\theta}^*$ such that

$$(4.2) \qquad B(\tilde{\theta}^*|F) = \min_{\tilde{\theta}} B(\tilde{\theta}|F)$$

then it is called the Bayes estimate of θ. Recalling that $W(F(\cdot|\tilde{\theta}))$ measures the discrepancy of a model $F(\cdot|\theta)$ from the true distribution $G(\cdot)$, we take $B(\tilde{\theta}^*|F)$ as a measure of the adequacy of a model $F(\cdot|\theta)$ associated with a prior distribution $P(\theta)$. That is, along the lines of previous sections, if we compare two alternative models, say, $F_1(\cdot|\theta)$ with $P_1(\theta)$ and $F_2(\cdot|\zeta)$ with $P_2(\zeta)$, then we decide to choose F_1 or F_2 according to whether or not $B(\tilde{\theta}^*|F_1) < B(\tilde{\zeta}^*|F_2)$.

In what follows let us be specific to a linear normal regression model for a vector random variable Y:

$$(4.3) \qquad F\colon \; Y \sim N\,(X\beta,\, \sigma^2 I_n)$$

where Y is n × 1, X is n × k, β is k × 1, and u is n × 1; the true

distribution of y is $N(\mu, \omega^2 I_n)$ with unknowns μ and ω^2. If we assume a diffuse prior for β and σ^2, the minimum attainable Bayes risk is evaluated as follows:

Lemma 4.1. Given a model F with a diffuse prior, the minimum attainable Bayes risk is

$$(4.4) \qquad B(\tilde{\beta}^*, \tilde{\sigma}^{2*}|F) = -\frac{2}{n} \log f\ (y|\hat{\beta}, \hat{\sigma}^2) + \log (\frac{n+k}{n-k-2}),$$

where $\hat{\beta}$ and $\hat{\sigma}^2$ are the ML estimates for β and σ^2, $\tilde{\beta}^*$ and $\tilde{\sigma}^{2*}$ are the Bayes estimates, and f is the density function of $N(X\beta, \sigma^2 I_n)$.

Let us make a comparison of two nested alternatives F_1 and F_2 given in (3.14). The Bayes decision rule, based on the magnitude of the minimum attainable Bayes risk, leads us to the following decision rule which is again described in terms of a familiar F-statistic.

Theorem 4.1. A decision rule based on the minimum attainable Bayes risk is equivalent to: choose F_1 if

$$(4.5) \qquad W < \frac{2(n-1)(n-p-q)}{(n+p)(n-p-q-2)} \quad ;$$

choose F_2 otherwise, where

$$(4.6) \qquad W = \frac{(n-p-q)(\hat{\sigma}_1^2 - \hat{\sigma}_2^2)}{q\hat{\sigma}_2^2}$$

is a F-statistic conventionally employed to test the hypothesis that $\beta_2 = 0$.

The proof is given in the Appendix.

We call the right-hand side of (4.5) the Bayes critical point, which tends to 2 asymptotically, increases with q, and decreases with p if n is moderately large. Limiting ourselves to the case of q = 1, we tabulate the numerical values of the square root of the Bayes critical point in Table 4.1 which is comparable to Tables 3.1 and 3.2.

Table 4.1

Bayes Critical Points for the Preliminary t-Test

n \ p	1	2	3	4	5
10	1.477	1.449	1.441	1.464	1.549
12	1.454	1.421	1.398	1.387	1.393
14	1.442	1.409	1.383	1.363	1.351
16	1.435	1.403	1.376	1.354	1.336
18	1.430	1.401	1.374	1.351	1.332
20	1.427	1.399	1.374	1.352	1.332
25	1.422	1.398	1.376	1.356	1.337
30	1.419	1.399	1.380	1.362	1.345
50	1.416	1.403	1.390	1.378	1.366
100	1.415	1.408	1.401	1.395	1.388
200	1.414	1.411	1.407	1.404	1.401
1000	1.414	1.414	1.413	1.412	1.411

It is interesting to note that the Bayes critical point varies quite little according to the changes in the values of n and p. Also, it is very close to the minimax regret critical point in Sawa and Hiromatsu [6].

5. Bias of Decision Rules

Now we return to Section 3 and reconsider the problem from the viewpoint of sampling theory. When we compare the two nested alternative models given in (3.14), our decision rule should be in principle based on the risk function given in Theorem 3.1. That is, we should choose F_1 if $R(F_1(\cdot|\theta_1)) < R(F_2(\cdot|\theta_2))$ and *vice versa*.

Lemma 5.1 If $\delta = \sigma_1^2 - \sigma_2^2 = 0\ (n^{-1})$, then

$$(5.1) \qquad R(F_1(\cdot|\theta_1)) - R(F_2(\cdot|\theta_2)) = \frac{\delta}{\sigma_2^2} - \frac{q\omega^2}{n\sigma_2^2} + 0\ (n^{-2}).$$

The proof is given in the Appendix. It should be recalled that when we derived the BIC the terms of $0(n^{-2})$ were neglected. It is, therefore, consistent that we evaluate the difference of risk only to order $0(n^{-1})$. The difference between the pseudo-variances, δ, is assumed to be $0(n^{-1})$. This assumption may seem to be somewhat uncomfortable. However, it may be justified by the fact that the model discrimination procedure would be unnecessary unless the difference between the two alternatives is as small as the reciprocal of the sample size.

Hence we can legitimately define a *correct decision rule* as follows: choose the model F_1 if $n\delta/\omega^2 < q$ and choose F_2 if $n\delta/\omega^2 < q$.

Based on the preceding consideration, we introduce the notion of *unbiasedness of a decision rule*: a decision rule is said to

be unbiased if the probability of choosing F_1 is greater than 1/2 when $n\delta/\omega^2 < q$ and less than 1/2 when $n\delta/\omega^2 > q$. If the probability decreases continuously with the increase of $n\delta/\omega^2$, the condition of unbiasedness is simply described as follows: the probability of choosing F_1 (or F_2) is 1/2 when $n\delta/\omega^2 = q$. Note that when $n\delta/\omega^2 = q$ we are indifferent to the two alternative models. If the above probability exceeds 1/2, then the decision rule is said to be biased toward a simpler model; if it falls below 1/2, then the decision rule is biased toward a more complex model.

All decision rules considered so far are based on whether or not an observed value of W, given by (4.6), exceeds a constant which changes with n, p, and q. Under the assumption that $Y \sim N(\mu, \omega^2 I_n)$, W is distributed as a doubly noncentral F with (q, n-p-q) degrees of freedom and the noncentrality parameters

$$(5.2) \qquad \delta_1 = \frac{n\delta}{\omega^2} = \frac{\mu' X_2^{*}(X_2^{*'} X_2^{*})^{-1} X_2^{*'} \mu}{\omega^2} \qquad \text{and}$$

$$(5.3) \qquad \delta_2 = \frac{\mu'[I - X_1(X_1'X_1)^{-1}X_1' - X_2^{*}(X_2^{*'}X_2^{*})^{-1}X_2^{*'}]\mu}{\omega^2}$$

where $X_2^{*} = X_2 - X_1(X_1'X_1)^{-1}X_1'X_2$. It would be worth noting here that a decision is correct if we decide to choose F_1 when the noncentrality parameter of the numerator is less than its degree of freedom and vice versa.

In Table 5.1 we tabulate the probability that W exceeds the BIC critical point when $n\delta/\omega^2 = q$, i.e., when F_1 and F_2 are indifferent. It can be observed from the Table that the BIC procedure is considerably biased toward a simpler model.

Table 5.1

Bias of the BIC Decision Rule

	noncentrality	n = 10	n = 20	n = 30	n = 40	n = 50
p = 2	.0	.696	.671	.664	.661	.659
	.1	.720	.697	.690	.687	.685
	.2	.742	.720	.714	.711	.709
	.3	.763	.742	.736	.733	.731
	.4	.781	.762	.756	.753	.752
	.5	.798	.780	.774	.772	.770
	.6	.814	.797	.791	.789	.788
	.7	.829	.812	.807	.805	.803
	.8	.842	.827	.822	.820	.818
	.9	.854	.840	.835	.833	.832
	1.0	.866	.852	.848	.846	.844
p = 3	.0	.738	.689	.675	.669	.666
	.1	.760	.715	.701	.695	.692
	.2	.781	.738	.725	.719	.715
	.3	.800	.759	.747	.741	.737
	.4	.817	.779	.767	.761	.758
	.5	.833	.797	.785	.779	.776
	.6	.848	.813	.802	.796	.793
	.7	.861	.828	.817	.812	.809
	.8	.873	.842	.832	.827	.824
	.9	.884	.855	.845	.840	.837
	1.0	.894	.866	.857	.852	.850

Each entry in the table is the probability that a doubly non-central F variate, with noncentrality parameters (δ_1, δ_2) and $(1, n - p - 1)$ degrees of freedom, falls below the BIC critical point when $\delta_1 = 1$. The noncentrality is $\delta_2/(n - p - 1)$, i.e., the normalized noncentrality parameter of the denominator in F, where δ_2 is given by (5.3).

Appendix

Proof of Lemma 3.1

The log likelihood function is

$$\text{(A.1)} \qquad \log f(y|\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}||y - X\beta||^2,$$

where $\theta' = (\beta', \sigma^2)$ and $||\cdot||$ stands for an Euclidean norm.

Differentiating it with respect to β and σ^2, we have

$$\text{(A.2)} \qquad \frac{\partial \log f(y|\theta)}{\partial \beta} = \frac{1}{\sigma^2} X'(y - X\beta),$$

$$\text{(A.3)} \qquad \frac{\partial \log f(y|\theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}||y - X\beta||^2.$$

Then

$$\text{(A.4)} \qquad E\left[\frac{\partial \log f(Y|\theta)}{\partial \beta}\right] = \frac{1}{\sigma^2} X'(\mu - X\beta)$$

$$\begin{aligned}\text{(A.5)} \qquad E\left[\frac{\partial \log f(Y|\theta)}{\partial \sigma^2}\right] &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} E\,||Y - X\beta||^2 \\ &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(E\,||Y - \mu||^2 + ||\mu - X\beta||^2) \\ &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\text{tr}\,\Omega + ||\mu - X\beta||^2).\end{aligned}$$

Equating (A.4) and (A.5) to zeroes and solving them yields the pseudo-true parameter values β_0 and σ_0^2 given, respectively, by (3.3) and (3.4).

Proof of Lemma 3.2

$$(A.6) \qquad E(\hat{\beta}) = (X'X)^{-1}X'\mu = \beta_0$$

$$(A.7) \qquad E(\hat{\sigma}^2) = \frac{1}{n} \operatorname{tr}\overline{P}_X(\mu\mu' + \omega^2 I_n)$$

$$= \frac{n-k}{n}\,\omega^2 + \frac{1}{n}\,\mu'\overline{P}_X\mu$$

where $\overline{P}_X = I - X(X'X)^{-1}X'$. Then

$$(A.8) \qquad \lim E(\hat{\sigma}^2) = \lim \sigma_0^2$$

Proof of Lemma 3.3

From (A.1) we have

$$(A.9) \qquad -\frac{2}{n} \log f(Y|\hat{\theta}) = \log (2\pi) + \log \hat{\sigma}^2 + \frac{1}{n\hat{\sigma}^2} \| Y - X\hat{\beta} \|^2$$

where Y is a vector random variable independent of $\hat{\theta}$. Taking expectation of (A.9) and substituting

$$(A.10) \qquad E[\| Y - X\hat{\beta} \|^2 | \hat{\beta}] = E\| Y - X\beta_0 \|^2 - 2\, E[(Y - X\beta_0)'X(\hat{\beta} - \beta_0)]$$

$$+ \| X(\hat{\beta} - \beta_0) \|^2$$

$$= n\sigma_0^2 - 2\mu'\overline{P}_X X(\hat{\beta} - \beta_0) + \| X(\hat{\beta} - \beta_0) \|^2$$

$$= n\sigma_0^2 + \| X(\hat{\beta} - \beta_0 \|^2$$

therein, we obtain (3.11).

Proof of Theorem 3.1

The risk function is

(A.11) $$R(F(\cdot|\theta)) = E[W(F(\cdot|\theta))]$$

$$= \log(2\pi) + \log(\sigma^2) + E[\log(\frac{\hat{\sigma}^2}{\sigma^2})] + E(\frac{\sigma^2}{\hat{\sigma}^2})$$

$$+ \frac{1}{n\sigma^2} E(\frac{\sigma^2}{\hat{\sigma}^2}) E\| X(\hat{\beta} - \beta) \|^2$$

where use is made of the independence of $\hat{\sigma}^2$ and $\hat{\beta}$, and the suffix 0 of σ_0^2 and β_0 is dropped. We have the following power series expansions:

(A.12) $$\log(\frac{\hat{\sigma}^2}{\sigma^2}) = \log(1 + \Delta) = \Delta - \frac{1}{2}\Delta^2 + \cdots$$

(A.13) $$\frac{\sigma^2}{\hat{\sigma}^2} = \frac{1}{1 + \Delta} = 1 - \Delta + \Delta^2 + \cdots$$

where

(A.14) $$\Delta = \frac{\hat{\sigma}^2 - \sigma^2}{\sigma^2}$$

Note that under the assumptions stated in the Theorem the expectations of higher order terms in the expansions are of order $0(n^{-2})$.

(A.15) $$\Delta = \frac{1}{n\sigma^2}[Y'\overline{P}_X Y - n\omega^2 - \mu'\overline{P}_X\mu]$$

$$= \frac{1}{n\sigma^2}[V'\overline{P}_X V + 2\mu'\overline{P}_X V] - \frac{\omega^2}{\sigma^2}$$

where $V = Y - \mu$. Under the assumptions in the Theorem

(A.16) $$E(V'\overline{P}_X V) = \omega^2 \mathrm{tr}\overline{P}_X = (n - k)\omega^2$$

(A.17) $$E(V'\overline{P}_X V)^2 = \omega^4[(\mathrm{tr}\overline{P}_X)^2 + 2\mathrm{tr}\overline{P}_X]$$

$$= [(n - k)^2 + 2(n - k)]\omega^4$$

(A.18) $$E(\mu'\overline{P}_X V) = E[V'\overline{P}_X V\mu'\overline{P}_X V] = 0$$

$$\text{(A.19)} \qquad E(\mu'\overline{P}_X V)^2 = \omega^2 \mu'\overline{P}_X \mu = n\omega^2(\sigma^2 - \omega^2)$$

Hence, rearranging the terms, we obtain

$$\text{(A.20)} \qquad E(\Delta) = -\frac{k}{n}\left(\frac{\omega^2}{\sigma^2}\right),$$

$$\text{(A.21)} \qquad E(\Delta^2) = \frac{4}{n}\left(\frac{\omega^2}{\sigma^2}\right) - \frac{2}{n}\left(\frac{\omega^2}{\sigma^2}\right)^2 + 0(n^{-2}).$$

Also, we have

$$\text{(A.22)} \qquad E\| X(\hat{\beta} - \beta) \|^2 = E\| X(X'X)^{-1}X'V \|^2 = \omega^2 \mathrm{tr} X(X'X)^{-1}X'$$

$$= k\omega^2$$

Therefore,

$$\text{(A.23)} \qquad E[\log(\frac{\sigma^2}{\hat{\sigma}^2})] + E(\frac{\sigma^2}{\hat{\sigma}^2}) = 1 + \frac{1}{2} E(\Delta^2) + 0(n^{-2})$$

$$= 1 + \frac{2}{n}\left(\frac{\omega^2}{\sigma^2}\right) - \frac{1}{n}\left(\frac{\omega^2}{\sigma^2}\right)^2 + 0(n^{-2})$$

$$\text{(A.24)} \qquad E\left(\frac{\sigma^2}{\hat{\sigma}^2}\right) E\| X(\hat{\beta} - \beta) \|^2 = k\omega^2 + 0(n^{-1})$$

Substituting (A.23) and (A.24) into (A.11), we finally obtain (3.12).

Proof of Theorem 3.2

From (A.12), (A.20) and (A.21)

$$\text{(A.25)} \qquad E(\log \hat{\sigma}^2) = \log \sigma^2 + E(\Delta) - \frac{1}{2} E(\Delta^2)$$

$$= \log \sigma^2 - \frac{k}{n}\left(\frac{\omega^2}{\sigma^2}\right) - \frac{2}{n}\left(\frac{\omega^2}{\sigma^2}\right) + \frac{1}{n}\left(\frac{\omega^2}{\sigma^2}\right)^2 + 0(n^{-2}).$$

Moreover, we have

$$\text{(A.26)} \qquad E\left(\frac{\hat{\omega}^2}{\hat{\sigma}^2}\right) = \frac{\omega^2}{\sigma_0^2} E\left(\frac{\sigma_0^2}{\hat{\sigma}^2}\right) = \frac{\omega^2}{\sigma_0^2}(1 + 0(n^{-1}))$$

and

$$\text{(A.27)} \qquad E\left(\frac{\hat{\omega}^2}{\hat{\sigma}^2}\right)^2 = \frac{\omega^4}{\sigma^4} E\left(\frac{\sigma_0^2}{\hat{\sigma}^2}\right)^2 = \frac{\omega^4}{\sigma^4}\,(1 + 0(n^{-1}))$$

Noting that

$$\text{(A.28)} \qquad -2 \log f(y|\hat{\theta}) = n \log (2\pi) + n \log \hat{\sigma}^2 + 1$$

and combining the above expectations, we obtain

$$\text{(A.29)} \qquad n\, E\, [BIC(F(\cdot|\theta))] = nR(F(\cdot|\theta)) + 0(n^{-1}).$$

Proof of Lemma 4.1

If we assume a linear normal regression model $Y \sim N(X\beta, \sigma^2 I)$ with diffuse prior for β and σ^2, the conditional posterior distribution of β, given σ^2, is $N(\hat{\beta}, \sigma^2(X'X)^{-1})$ where $\hat{\beta} = (X'X)^{-1}X'y$ is the maximum likelihood estimate, and also the marginal prior distribution for σ^2 is the inverse gamma distribution with the density function

$$\text{(A.30)} \qquad \frac{2}{\Gamma(\nu/2)} \left(\frac{\nu s^2}{2}\right)^{\nu/2} \frac{1}{\sigma^{\nu+1}} \exp\left(-\frac{\nu s^2}{2\sigma^2}\right)$$

where $\nu = n - k$ and $s^2 = n\hat{\sigma}^2/(n - k)$. The proof is given by Zellner [8]. The conditional expectation of $\| X(\tilde{\beta} - \beta) \|^2$ with respect to the posterior distribution is

$$\begin{aligned}\text{(A.31)} \qquad E_{\beta|y,\sigma} \| X(\tilde{\beta} - \beta) \|^2 &= E_{\beta|y,\sigma} \| X(\beta - \hat{\beta})\|^2 + \| X(\hat{\beta} - \tilde{\beta}) \|^2 \\ &\geq E_{\beta|y,\sigma} \| X(\beta - \hat{\beta}) \|^2 \\ &= k\sigma^2\end{aligned}$$

where the lower bound is attainable when $\tilde{\beta} = \hat{\beta}$; i.e., the Bayes estimate $\tilde{\beta}^*$ of β is nothing but the ML estimate. A straightforward integration

yields

$$\text{(A.32)}\qquad E_{\sigma^2|y}(\sigma^2) = \frac{\nu}{\nu - 2}\, s^2 = \frac{n - k}{n - k - 2}\, s^2$$

as long as $\nu > 2$. Hence

$$\text{(A.33)}\qquad E_{\beta,\sigma^2|y}[W(F(\cdot|\theta))] \geq \log\,(2\pi) + \log\,\tilde{\sigma}^2 + \frac{n + k}{n - k - 2}\,(1 + \frac{k}{n})\,\frac{s^2}{\tilde{\sigma}^2}$$

The Bayes estimate of σ^2 is $\tilde{\sigma}^2$ that minimizes the right-hand side of the above inequality; i.e.

$$\text{(A.34)}\qquad \tilde{\sigma}^{*2} = \frac{n + k}{n - k - 2}\,\hat{\sigma}^2$$

where $\hat{\sigma}^2$ is the ML estimate of σ^2. Substituting this into the right-hand side of (A.33), the minimum attainable Bayes risk is evaluated as follows:

$$\begin{aligned}\text{(A.35)}\qquad B(\tilde{\beta}^*, \tilde{\sigma}^{*2}|F) &= \log\, 2\pi + \log\,\tilde{\sigma}^{*2} + 1 \\ &= \log\, 2\pi + \log\,\hat{\sigma}^2 + 1 + \log\,(\frac{n + k}{n - k - 2}) \\ &= -\,\frac{2}{n}\log\, f(y|\hat{\theta}) + \log(\frac{n + k}{n - k - 2}).\end{aligned}$$

<u>Proof of Theorem 4.1</u>

Let B_1 and B_2 be the minimum attainable Bayes risks, respectively, for F_1 and F_2 with diffuse prior for parameters. The difference between B_1 and B_2 is

$$\text{(A.36)}\qquad B_1 - B_2 = \log\,(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}) + \log\,[\frac{(n + p)(n - p - q - 2)}{(n + p + q)(n - p - 2)}]$$

If this is negative, we should choose F_1, and *vice versa*. By the monotonicity of the logarithm transformation, $B_1 - B_2 < 0$ is equivalent to

$$\text{(A.37)} \qquad \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} < \frac{(n + p + q)(n - p - 2)}{(n + p)(n - p - q - 2)}$$

which is again equivalent to (4.5).

Proof of Lemma 5.1

$$\text{(A.38)} \qquad R(F_1(\cdot|\theta)) - R(F_2(\cdot|\theta)) = \log\left(\frac{\sigma_1^2}{\sigma_2^2}\right) + \frac{p + 2}{n}\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right)\omega^2$$

$$- \frac{q}{n}\frac{\omega^2}{\sigma_2^2} + \frac{1}{n}\left(\frac{1}{\sigma_2^4} - \frac{1}{\sigma_1^4}\right)\omega^4 + O(n^{-2}).$$

If we assume that

$$\text{(A.39)} \qquad \delta = \sigma_1^2 - \sigma_2^2 = O(n^{-1}),$$

we have an expansion

$$\text{(A.40)} \qquad \log\left(\frac{\sigma_1^2}{\sigma_2^2}\right) = \log\left(1 + \frac{\delta}{\sigma_2^2}\right) = \frac{\delta}{\sigma_2^2} + O(n^{-2}).$$

Also, it follows that the second and third terms on the right-hand side of (A.38) are of order $O(n^{-2})$. Hence, if we neglect the terms of order $O(n^{-2})$, we can assert that $R(F_1(\cdot|\theta)) < R(F_2(\cdot|\theta))$ if and only if

$$\text{(A.41)} \qquad \frac{n\delta}{\omega^2} < q$$

and *vice versa*.

REFERENCES

[1] Akaike, H. (1972) "Information Theory and an Extension of the Maximum Likelihood Principle," in *Proc. 2nd Int. Symp. on Information Theory*, pp. 267-281.

[2] Akaike, H. (1974) "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, Vol. AC-19, No. 6, pp. 716-723.

[3] Kullback, S. (1959) *Information Theory and Statistics*, New York, John Wiley and Sons.

[4] Mallows, C. L. (1973) "Some Comments on Cp", *Technometrics*, Vol. 15, pp. 661-675.

[5] Rao, C. R. (1973) *Linear Statistical Inference and Its Application*, 2nd ed., New York, John Wiley and Sons.

[6] Sawa, T. and T. Hiromatsu (1973) "Minimax Regret Significance Points for a Preliminary Test in Regression Analysis," *Econometrica*, Vol. 41, pp. 1093-1101.

[7] Takeuchi, Kei (1976) "On Sampling Distribution of Criteria for Selection of Independent Variables Related with C_p-Statistic," mimeographed, presented at International Biometric Conference at Boston, August 1976.

[8] Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics*, New York, John Wiley and Sons.