SELECTION OF ECONOMETRIC MODELS

BY THE INFORMATION CRITERION

Gregory C. Chow*

Econometric Research Program
Research Memorandum No. 239

March 1979

Abstract

This paper provides an exposition of the information criterion for model selection, suggests an estimate of it, presents several applications to the selection of models of simultaneous equations, compares the information criterion with the posterior probability criterion, and indicates several related econometric problems for research.

* The author acknowledges financial support from the National Science Foundation.

Econometric Research Program
Princeton University
207 Dickinson Hall
Princeton, New Jersey

SELECTION OF ECONOMETRIC MODELS
BY THE INFORMATION CRITERION

Gregory C. Chow

1. Introduction

This paper provides an exposition of the information criterion, presents an estimate of it, discusses some of its applications to the selection of econometric models, and compares it with the posterior probability criterion for model selection. Recently, Sawa (1978) has given an excellent exposition of the information criterion and obtained an estimate of it for the selection of linear regression models. The present exposition is different from and complementary to Sawa's. The estimate of information suggested here is of more general applicability while it reduces to Sawa's in the regression situation. This paper is also concerned with econometric applications other than the selection of linear regression models.

In order to introduce the information criterion, we will first consider, in Section 2, the selection of linear regression models for the purpose of prediction, using the expected squared prediction error as the criterion. We will present in Section 3 the concept of mean information, derive the estimate of it as originally given by Akaike (1973, 1974) and suggest an improvement of Akaike's estimate. Section 4 deals with several problems arising from the selection of simultaneous-equation models using the information criterion. In Section 5, we compare the information criterion with the posterior-odds ratio advocated by some Bayesian statisticians for the purpose of model selection, and point out their differences. The paper ends with some brief concluding remarks in Section 6.

2. Selection of Regression Models by Prediction

Consider the choice between the regression model

$$(2.1)\quad Y = X\beta + u = X_1\beta_1 + X_2\beta_2 + u$$

and the linear model using X_1 alone as the explanatory variables, where X_1 is n by k_1 and X_2 is n by k_2, u being normal with covariance matrix $I_n\sigma^2$. A standard treatment is to test the null hypothesis $\beta_2 = 0$ using the F ratio, but this does not solve the problem as it begs the question of what level of significance to use. A more satisfactory solution is to choose the model which is estimated to have smaller prediction errors. Specifically, let n new observations be

$$(2.2)\quad \tilde{Y} = \tilde{X}\beta + \tilde{u}$$

under the assumption that (2.1) is the true regression model, and let the model selection criterion be the expected sum of squared prediction errors. One can derive a rule for choosing between the two models, as given by Mallows (1973).

Using the small model with X_1 alone and denoting the corresponding maximum likelihood estimate of β by $\hat{\beta}_I$ (consisting of $(X_1'X_1)^{-1}X_1'Y$ and 0), one easily finds

$$(2.3)\quad E(\hat{\beta}_I-\beta)(\hat{\beta}_I-\beta)' = \begin{bmatrix} (X_1'X_1)^{-1}X_1'X_2\beta_2\beta_2'X_2'X_1(X_1'X_1)^{-1}+(X_1'X_1)^{-1}\sigma^2 & -(X_1'X_1)^{-1}X_1'X_2\beta_2\beta_2' \\ -\beta_2\beta_2'X_2'X_1(X_1'X_1)^{-1} & \beta_2\beta_2' \end{bmatrix}.$$

The expected sum of squared prediction errors is

$$(2.4)\quad E(\tilde{X}\hat{\beta}_I-\tilde{Y})'(\tilde{X}\hat{\beta}_I-\tilde{Y}) = E(\hat{\beta}_I-\beta)'\tilde{X}'\tilde{X}(\hat{\beta}_I-\beta) + E\tilde{u}'\tilde{u}$$

$$= \mathrm{tr}[\tilde{X}'\tilde{X}E(\hat{\beta}_I-\beta)(\hat{\beta}_I-\beta)'] + n\sigma^2$$

$$= k_1\sigma^2 + \beta_2'X_2'[I-X_1(X_1'X_1)^{-1}X_1]X_2\beta_2 + n\sigma^2$$

where the last line has utilized (2.3) and the reasonable assumption $\tilde{X}'\tilde{X} = X'X$ made to provide a standard for comparing the two models. Using the large model (2.1) and denoting the maximum likelihood estimate of β by $\hat{\beta} = (X'X)^{-1}X'Y$, we have

$$(2.5)\quad E(\tilde{X}\hat{\beta}-\tilde{Y})'(\tilde{X}\hat{\beta}-\tilde{Y}) = E(\hat{\beta}-\beta)'\tilde{X}'\tilde{X}(\hat{\beta}-\beta) + E\tilde{u}'\tilde{u} = (k_1+k_2)\sigma^2 + n\sigma^2$$

Comparing (2.4) and (2.5), we find that the small model should be used if and only if

$$(2.6)\quad \beta_2'X_2'[I-X_1(X_1'X_1)^{-1}X_1']X_2\beta_2 \equiv \beta_2'X_{2.1}'X_{2.1}\beta_2 < k_2\sigma^2$$

where $X_{2.1} = [I-X_1(X_1'X_1)^{-1}X_1']X_2$ denotes the matrix of residuals of the regression of X_2 on X_1.

Since we do not know β_2 and σ^2, we have to estimate them for the application of (2.6). Given that $\hat{\beta}_2$ has mean β_2 and covariance matrix $(X_{2.1}'X_{2.1})^{-1}\sigma^2$, we have

$$(2.7)\quad E(\hat{\beta}_2-\beta_2)'(X_{2.1}'X_{2.1})(\hat{\beta}_2-\beta_2)/\sigma^2 = (E\hat{\beta}_2'X_{2.1}'X_{2.1}\hat{\beta}_2-\beta_2'X_{2.1}'X_{2.1}\beta_2)/\sigma^2 = k_2$$

$$\text{or}\qquad E\hat{\beta}_2X_{2.1}'X_{2.1}\hat{\beta}_2 = \beta_2'X_{2.1}'X_{2.1}\beta_2 + k_2\sigma^2$$

The selection criterion (2.6) is equivalent to

$$(2.8)\quad \beta_2'X_{2.1}'X_{2.1}\beta_2 + k_2\sigma^2 < 2k_2\sigma^2 .$$

If the left-hand side of (2.8) is replaced by the unbiased estimate from (2.7) and σ^2 on the right-hand side is replaced by the unbiased estimate s^2 , we obtain

$$(2.9)\quad \hat{\beta}_2' X_{2.1}' X_{2.1} \hat{\beta}_2 < 2k_2 s^2 \equiv 2k_2 (Y-X\hat{\beta})'(Y-X\hat{\beta})/(n-k_1-k_2) \ .$$

as the condition for selecting the small model. In the language of testing hypothesis, this rule amounts to setting the critical value for the F statistic $\hat{\beta}_2' X_{2.1}' X_{2.1} \hat{\beta}_2 / k_2 s^2$ under the null hypothesis $\beta_2 = 0$ to 2. We observe from (2.4) and (2.5) that omitting the variables X_2 might yield a better model for prediction even when $\beta_2 \neq 0$ because $E(\hat{\beta}_I - \beta)'\tilde{X}'\tilde{X}(\hat{\beta}_I - \beta)$ in (2.4) might be smaller than $E(\hat{\beta}-\beta)'\tilde{X}'\tilde{X}(\hat{\beta}-\beta)$ in (2.5) . In other words, a misspecification might lead to better prediction as it reduces the covariance matrix $E(\hat{\beta}-\beta)(\hat{\beta}-\beta)'$ of estimation errors.

The criterion of expected squared prediction error applies to the selection of non-nested regression models. Let the two models have regression functions $X_1\beta_1$ and $X_2\beta_2$ respectively, where X_1 and X_2 are disjointed. Under the assumption that (2.1) is the true model, we have an expression analogous to (2.4) for the expected sum of squared prediction errors resulting from using the second model with X_2 alone as the explanatory variables. Comparing these two expressions, we will select the first model if

$$(2.10)\quad k_1\sigma^2 + \beta_2' X_{2.1}' X_{2.1} \beta_2 < k_2\sigma^2 + \beta_1' X_{1.2}' X_{1.2} \beta_1$$

Replacing both sides of (2.10) by the unbiased estimates from (2.7) and rearranging terms we obtain

$$(2.11)\quad \hat{\beta}_2' X_{2.1}' X_{2.1} \hat{\beta}_2 - \hat{\beta}_1 X_{1.2}' X_{1.2} \hat{\beta}_1 < 2(k_2-k_1)s^2$$

as the condition for choosing the first model with X_1 as the explanatory variables.

The information criterion for model selection suggested by Akaike (1973, 1974) can be viewed as an extension of the above prediction criterion. Instead of using the expected squared prediction error of a future observation, Akaike has adopted the mean information $E[\log g(\tilde{y}) - \log f(\tilde{y})]$ for discrimination between the density function $g(\cdot)$ of the true model and the density function $f(\cdot)$ of the model used for prediction. In the nested regression example, the log-likelihood of the new observation vector $\tilde{Y} = \tilde{X}\beta + \tilde{u}$ for the true model is

$$(2.12)\quad \sum_{1}^{n} \log g(\tilde{y}_i) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2}(\tilde{Y}-\tilde{X}\beta)'(\tilde{Y}-\tilde{X}\beta)/\sigma^2$$

and the log-likelihood of $\tilde{Y}$ according to the estimated small model is

$$(2.13)\quad \sum_{1}^{n} \log f(\tilde{y}_i) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \hat{\sigma}_I^2 - \frac{1}{2}(\tilde{y}-\tilde{X}\hat{\beta}_I)'(\tilde{y}-\tilde{X}\hat{\beta}_I)/\hat{\sigma}_I^2$$

where $\hat{\sigma}_I^2 = (y-X\hat{\beta}_I)'(y-X\hat{\beta}_I)/n$. The mean information criterion for given $\hat{\beta}_I$ and $\hat{\sigma}_I$ is

$$(2.14)\quad E_{\tilde{Y}}\sum_{1}^{n}[\log g(\tilde{y}_i)-\log f(\tilde{y}_i)] = -\frac{n}{2}(\log \sigma^2-\log\hat{\sigma}_I^2) - \frac{n}{2} - \frac{1}{2}E_{\tilde{Y}}(\tilde{Y}-\tilde{X}\hat{\beta}_I)'(\tilde{Y}-\tilde{X}\hat{\beta}_I)/\hat{\sigma}_I^2$$

Note that the criterion (2.14) measures the goodness of the estimated model not only by the expected sum of squared prediction errors $E_{\tilde{Y}}(\tilde{Y}-\tilde{X}\hat{\beta}_I)'(\tilde{Y}-\tilde{X}\hat{\beta}_I)$, but also by the difference between $\log \sigma^2$ and $\log \hat{\sigma}_I^2$ due to the error in $\hat{\sigma}_I^2$ as an estimate of the residual variance σ^2 of the regression residual. In (2.14), the estimated parameters $\hat{\beta}_I$ and $\hat{\sigma}_I^2$ of the small model are treated as fixed for the purpose of defining the information measure; only $\tilde{Y}$ is treated as random when the expectation is evaluated. As the next step, one should take the expectation $E_{\hat{\beta}_I,\hat{\sigma}_I^2}$ of (2.14), treating $\hat{\beta}_I$ and $\hat{\sigma}_I^2$ as random variables, in order to evaluate the estimated model by allowing for the sampling errors of its parameters. Both steps were incorporated in (2.4) when the expected squared error of prediction was used to evaluate this model. The expectation $E(\cdot)$ in (2.4) could have been replaced by $E_{\hat{\beta}_I}E_{\tilde{Y}}(\cdot\,|\,\hat{\beta}_I)$ to make clear these two steps.

3. Expected Information and its Estimation

Let $f(.|\theta)$ be used to predict n new independent observations $\tilde{Y} = (\tilde{y}_1,\dots,\tilde{y}_n)$ generated by the true density $g(.)$. The information measure of the discrepancy of the predictions is

$$(3.1) \qquad I\left[g;f(.|\theta)\right] = E \sum_{i=1}^{n} \left[\log g(\tilde{y}_i) - \log f(\tilde{y}_i|\theta)\right]$$

where the expectation is evaluated by the true density $g(.)$, and $\tilde{y}_i$ may denote a vector. This measure was suggested by Kullback and Leibler (1951), used by Savage (1954), and studied more fully by Kullback (1959). Akaike (1973, 1974) advocated the use of this measure and provided an estimate of it for model selection. It can be shown (Rao, 1973, pp. 58-59) that $I[g;f(\cdot|\theta)]$ is non-negative and is equal to zero if and only if $f(x|\theta) = g(x)$ almost everywhere.

If two functions $f_a(\cdot|\theta_a)$ and $f_b(\cdot|\theta_b)$ are being considered, the one with a smaller value for the mean information $I[g;f_i(\cdot|\theta_i)]$ will be chosen. If θ_i is estimated by some estimator $\hat{\theta}_i$, to choose between two estimated density functions $f_i(\cdot|\hat{\theta}_i)$ (i=a,b), one allows for the sampling errors of $\hat{\theta}_i$ and selects the one having a smaller value for the expected mean information $E_{\hat{\theta}_i} I[g;f(\cdot|\hat{\theta}_i)]$.

To obtain an estimate of $E_{\hat{\theta}} I[g;f(\cdot|\hat{\theta})]$, Akaike (1973) assumes that $g(\cdot) = f(\cdot|\theta^o)$, with $\theta = (\theta_1,\theta_2)$, and that the approximating $f(\cdot|\theta_1,0)$ is formed by restricting a subset θ_2 of the parameters to be zero. Although Akaike's estimate has been widely used, its derivation in Akaike (1973, pp. 273-276; 1974) seems to be obscure. We hope to point out the assumptions used in deriving his estimate and suggest an improvement of it. First, for any given estimate $\hat{\theta}$ of θ, approximate $I[g;f(\cdot|\hat{\theta})]$ by expanding $\sum_{i=1}^{n} \log f(\tilde{y}_i|\hat{\theta}) \equiv \log L(\tilde{Y}|\hat{\theta})$ in a second-order Taylor series about θ^o.

$$(3.2) \quad \log L(\tilde{Y}|\hat{\theta}) = \log L(\tilde{Y}|\theta^o) + (\hat{\theta}-\theta^o)'\frac{\partial \log L(\tilde{Y}|\theta^o)}{\partial\theta} + \frac{1}{2}(\hat{\theta}-\theta^o)'\frac{\partial^2 \log L(\tilde{Y}|\tilde{\theta})}{\partial\theta\partial\theta'}(\hat{\theta}-\theta^o)$$

where $\hat{\theta} \leq \tilde{\theta} \leq \theta^o$. Substitution of (3.2) into (3.1) gives

$$(3.3) \quad I[g;f(\cdot|\hat{\theta})] = E_{\tilde{Y}}[\log L(\tilde{Y}|\theta^o) - \log L(\tilde{Y}|\hat{\theta})]$$

$$= -(\hat{\theta}-\theta^o)'E_{\tilde{Y}}(\frac{\partial \log L(\tilde{Y}|\theta^o)}{\partial\theta}) - \frac{1}{2}(\hat{\theta}-\theta^o)'E_{\tilde{Y}}(\frac{\partial^2 \log L(\tilde{Y}|\tilde{\theta})}{\partial\theta\partial\theta'})(\hat{\theta}-\theta^o)$$

$$= \frac{1}{2}(\hat{\theta}-\theta^o)'J(\tilde{\theta},\theta^o)(\hat{\theta}-\theta^o)$$

where we have observed

$$E_{\tilde{Y}}(\frac{\partial \log L(\tilde{Y}|\theta^o)}{\partial\theta}) = \int \frac{1}{L(X|\theta^o)} \frac{\partial \cancel{\log} L(X|\theta^o)}{\partial\theta} \cdot L(X|\theta^o)dX = 0$$

and have defined

$$(3.4) \quad J(\tilde{\theta},\theta^o) \equiv -\int \frac{\partial^2 \log L(X|\tilde{\theta})}{\partial\theta\partial\theta'} L(X|\theta^o)dX \ .$$

Note that $J(\theta^o,\theta^o)$ is Fisher's information matrix for the density function $L(\cdot|\theta^o)$. In the approximation used by Akaike, $J(\tilde{\theta},\theta^o)$ in (3.3) was replaced by $J(\theta^o,\theta^o)$, but we will retain $J(\tilde{\theta},\theta^o)$ in (3.3).

Second, we define the best approximate model $f(\cdot|\theta_1^*,0)$ for predicting $\tilde{Y}$ by

$$(3.5) \quad I[g;f(\cdot|\theta_1^*,0)] \leq I[g;f(\cdot|\theta_1,0)] \quad \text{or}$$

$$E_{\tilde{Y}}[\log L(\tilde{Y}|\theta_1^*,0)] \geq E_{\tilde{Y}}[\log L(\tilde{Y}|\theta_1,0)]$$

In the terminology of Sawa (1978), θ_1^* is the pseudo-true parameter of the pseudo-true model $f(\cdot|\theta_1^*,0)$. Using the quadratic approximation (3.3), we find θ_1^* by minimizing

$$(3.6) \quad I[g;f(\cdot|\theta_1,0)] = \frac{1}{2}[(\theta_1-\theta_1^o)' \quad -\theta_2^{o\prime}] \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix} \begin{bmatrix} \theta_1-\theta_1^o \\ -\theta_2^o \end{bmatrix}$$

where the arguments of $J(\tilde{\theta},\theta^o)$ are omitted. Differentiation of (3.6) yields

$$(3.7) \quad \theta_1^* = \theta_1^o + J_{11}^{-1} J_{12} \theta_2^o \ .$$

In the regression example of (2.3), (3.7) would imply

$$(3.8)\qquad \beta_1^* = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 \ .$$

As the third step, we evaluate $I[g;f(\cdot|\hat{\theta}_1^*,0]$ where $\hat{\theta}_1^*$ is an estimate of θ_1^* , using equations (3.3), (3.7) and the identity

$$(\hat{\theta}_1^*,0) - (\theta_1^o,\theta_2^o) = (\hat{\theta}_1^*,0) - (\theta_1^*,0) + (\theta_1^*,0) - (\theta_1^o,\theta_2^o) \ ,$$

$$(3.9)\qquad I[g;f(\cdot|\hat{\theta}_1^*,0)] = \tfrac{1}{2}[(\hat{\theta}_1^*-\theta_1^o)' \quad -\theta_2^o] \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix} \begin{bmatrix} \hat{\theta}_1^*-\theta_1^o \\ -\theta_2^o \end{bmatrix}$$

$$= \tfrac{1}{2}[\hat{\theta}_1^*-\theta_1^*]'J_{11}[\hat{\theta}_1^*-\theta_1^*] + \tfrac{1}{2}[(\theta_1^*-\theta_1^o)' \quad -\theta_2^{o\prime}] \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix} \begin{bmatrix} \theta_1^*-\theta_1^o \\ -\theta_2^o \end{bmatrix}$$

$$= \tfrac{1}{2}[\hat{\theta}_1^*-\theta_1^*]'J_{11}[\hat{\theta}_1^*-\theta_1^*] + \tfrac{1}{2}\theta_2^{o\prime}[J_{22}-J_{21}J_{11}^{-1}J_{12}]\theta_2^o$$

where the cross-product term on the second line vanishes because of (3.7), and the second term on the last line is also due to (3.7). Note that the first term (on the last line) of (3.9) measures the contribution of the <u>sampling error</u> $\hat{\theta}_1^*-\theta_1^*$, and the second term measures the contribution of the <u>specification error</u> $\theta^*-\theta^o = (\theta_1^*,0) - (\theta_1^o,\theta_2^o)$, to the discrepancy $I[g;f(\cdot|\hat{\theta}_1^*,0)]$ between the estimated model $f(\cdot|\hat{\theta}_1^*,0)$ and the true density $g(\cdot) = f(\cdot|\theta_1^o,\theta_2^o)$. If $\hat{\theta}_1^*$ is the maximum likelihood estimator, the expectations of these two terms will correspond to the first two terms on the last line of (2.4) in our regression example. (The last term $n\sigma^2$ in (2.4) is not applicable because the true model still has a prediction error and we are measuring the discrepency between the predictive abilities of the true model and the estimated, approximate model.)

Fourth, we evaluate the second term using the quadratic approximation (3.3),

$$(3.10)\qquad \frac{1}{2}(\theta^*-\theta^o)'J(\theta^*-\theta^o) = I[g;f(.|\theta^*)] = E_{\tilde{Y}}\log L(\tilde{Y}|\theta^o) - E_{\tilde{Y}}\log L(\tilde{Y}|\theta^*)$$

Given a sample $Y = (y_1,\ldots,y_n)$ of n independent observations, we estimate $E_{\tilde{Y}}\log L(\tilde{Y}|\theta^*)$ by $\log L(Y|\theta^*)$ and expand the latter about the maximum likelihood estimate $(\hat{\theta}_1^*,0)$,

$$(3.11)\qquad \begin{aligned} E_{\tilde{Y}}\log L(\tilde{Y}|\theta^*) &\simeq \log L(Y,\theta^*) \\ &= \log L(Y;\hat{\theta}_1^*,0) + \frac{1}{2}(\hat{\theta}_1^*-\theta_1^*)\frac{\partial^2\log L(Y;\hat{\theta}_1^*,0)}{\partial\theta_1\partial\theta_1'}(\hat{\theta}_1^*-\theta_1^*) \\ &= \log L(Y;\hat{\theta}_1^*,0) - \frac{1}{2}(\hat{\theta}_1^*-\theta_1^*)'J_{11}(\theta^*,\theta^o)(\hat{\theta}_1^*-\theta_1^*) \end{aligned}$$

where, on the second line, we have observed that $\partial\log L(Y;\hat{\theta}_1^*,0)/\partial\theta_1$ is zero and, on the last line, we have replaced $\partial^2\log L(Y;\hat{\theta}_1^*,0)/\partial\theta_1\partial\theta_1'$ by the expectation $-J_{11}(\theta^*,\theta^o)$, noting that the data y_i are generated by the true density $f(y|\theta^o)$ and that the maximum likelihood estimator $\hat{\theta}_1^*$ converges to θ_1^*. We will discuss the convergence of $\hat{\theta}_1^*$ to θ_1^* below.

Combining (3.9), (3.10), and (3.11), we write

$$(3.12)\qquad \begin{aligned} -I[g;f(.|\hat{\theta}_1^*,0)] &= \log L(Y;\hat{\theta}_1^*,0) - (\hat{\theta}_1^*-\theta_1^*)'J_{11}(\theta^*,\theta^o)(\hat{\theta}_1^*-\theta_1^*) \\ &\quad - E_{\tilde{Y}}\log L(\tilde{Y}|\theta^o). \end{aligned}$$

Note that the first argument of J_{11} in (3.9) is $\tilde{\theta}$ which is defined in (3.2) to satisfy $\hat{\theta} \le \tilde{\theta} \le \theta^o$; it is replaced by θ^* in (3.12). If all models to be compared are specified by setting a subset of the parameters θ in the true density $f(.|\theta)$ equal to zero, the last term $-E_{\tilde{Y}}\log L(\tilde{Y}|\theta^o)$ in (3.12) is common to all models and does not have to be estimated. The rule is to choose the model with the largest $-E_{\hat{\theta}_1^*}I[g;f(.|\hat{\theta}_1^*,0)]$ which, by (3.12), can be estimated by

$$(3.13)\qquad \log L(Y;\hat{\theta}_1^*,0) - E_{\hat{\theta}_1^*}(\hat{\theta}_1^*-\theta_1^*)'[J_{11}(\theta^*,\theta^o)](\hat{\theta}_1^*-\theta_1^*) - E_{\tilde{Y}}\log L(\tilde{Y}|\theta^o)$$

where the constant term $-E_{\tilde{Y}}\log L(\tilde{Y}|\theta^o)$ can be ignored. Let θ_1 contain k parameters. Akaike (1973, p. 275) writes, "it can be shown that ... $N||_k\hat{\theta}-_k\theta||_c^2$ [or $(\hat{\theta}_1^*-\theta_1^*)'[J_{11}(\theta^o,\theta^o)](\hat{\theta}_1^*-\theta_1^*)$ in our notations] is asymptotically distributed as a chi-square variable with k degrees of freedom." If this claim is valid, and if we replace $J_{11}(\theta^*,\theta^o)$ in (3.13) by $J_{11}(\theta^o,\theta^o)$ then the second term of (3.13) will equal k . The selection criterion amounts to the maximum value of the log-likelihood minus the number k of parameters. It is known as Akaike's Information Criterion.

However, the above claim of Akaike may be questioned. If the observations were generated by the approximate model $f(\cdot|\theta_1^*,0)$, then $(\hat{\theta}_1^*-\theta_1^*)'[J_{11}(\theta^*,\theta^*)](\hat{\theta}_1^*-\theta_1^*)$ would be asymptotically $\chi^2(k)$. In the present situation, the observations are generated by the model $f(\cdot|\theta^o)$, but the maximum likelihood estimate $\hat{\theta}_1^*$ is based on the approximate (incorrect) model $f(\cdot|\theta^*,0)$. Akaike's information criterion presumes that $[J_{11}(\theta^*,\theta^o)]^{-1}$ is a good approximation to the covariance matrix of the estimator $\hat{\theta}_1^*$ which is computed from the approximate density function $f(\cdot|\theta_1,0)$. Let us therefore derive the asymptotic distribution of $\hat{\theta}_1^*$ by a Taylor expansion of $\partial\log L(Y;\hat{\theta}_1^*,0)/\partial\theta_1$ about θ_1^* :

$$(3.14)\quad \frac{\partial\log L(Y;\hat{\theta}_1^*,0)}{\partial\theta_1} = \frac{\partial\log L(Y;\theta_1^*,0)}{\partial\theta_1} + [\frac{\partial^2\log L(Y;\theta_1^*,0)}{\partial\theta_1\partial\theta_1'} + o(n)](\hat{\theta}_1^*-\theta_1^*)=0$$

$\partial\log L(Y;\theta_1^*,0)/\partial\theta_1$ has expectation zero since θ_1^* is derived by maximizing (differentiating) $E\log L(Y;\theta_1,0)$ with respect to θ_1. It is the sum of n independent $\log f(y_i|\theta_1^*,0)/\partial\theta_1$ and is therefore asymptotically normal by the central limit theorem. Its covariance matrix will be denoted by $\mathrm{Cov}[\partial\log L(Y;\theta^*)/\partial\theta_1]$. The matrix coefficient of $(\hat{\theta}_1^*-\theta_1^*)$ in (3.14) has expectation $E[\partial^2\log L(Y;\theta^*)/\partial\theta_1\partial\theta_1'] = -J_{11}(\theta^*,\theta^o)$. Therefore, $(\hat{\theta}_1^*-\theta_1^*)$ as a solution to (3.14) is asymptotically normal with zero mean and covariance matrix

$$(3.15) \qquad \Phi = [J_{11}(\theta^*,\theta^o)]^{-1}\mathrm{Cov}[\frac{\partial \log L(Y;\theta^*)}{\partial\theta_1}][J_{11}(\theta^*,\theta^o)]^{-1}.$$

The result (3.15) can be deduced from Lemma 1 of Silvey (1959, p. 394) who considers the distribution of the maximum likelihood estimator subject to more general restrictions of the form $h(\theta) = 0$ under the condition that the restrictions are incorrect. Substituting (3.15) for $E(\hat\theta_1^*-\theta_1^*)(\hat\theta_1^*-\theta_1^*)$ in (3.13) we obtain the following criterion for model selection

$$(3.16) \quad \log L\ (Y;\hat\theta_1^*,0) - \mathrm{tr}\{\mathrm{Cov}[\frac{\partial \log\ L(Y|\theta^*)}{\partial\theta_1}][J_{11}(\theta^*,\theta^o)]^{-1}\} - E_{\tilde Y}\log\ L(\tilde Y|\theta^o)$$

It is well-known that if the true density were $f(\cdot|\theta^*)$, Fisher's information matrix $J_{11}(\theta^*,\theta^*)$ equals $\mathrm{Cov}_{\theta^*}[\frac{\partial\log\ L(Y|\theta^*)}{\partial\theta_1}]$. However, if the observations are generated by the true density $f(\cdot|\theta^o)$, $J_{11}(\theta^*,\theta^o)$ is in general not equal to $\mathrm{Cov}_{\theta^o}[\frac{\partial\log\ L(Y|\theta^*)}{\partial\theta_1}]$. In order to estimate these two matrices for the application of (3.16) to model selection, we propose to use respectively

$$(3.17) \qquad -\frac{\partial^2\log L(Y;\hat\theta_1^*,0)}{\partial\theta_1\partial\theta_1'} \quad \text{and} \quad E_{\theta^o}\left[\frac{\partial\log L(Y;\theta^*)}{\partial\theta_1}\cdot\frac{\partial\log L(Y;\theta^*)}{\partial\theta_1'}\right]$$

where the parameters θ^o and θ^* of the last expectation will be replaced by their maximum likelihood estimates. Berndt, Hall, Hall and Hausman (1974) have proposed to use the inverse of $n^{-1}\sum_i \frac{\partial\log\ f(y_i|\hat\theta)}{\partial\theta}\cdot\frac{\partial\log\ f(y_i|\hat\theta)}{\partial\theta'}$ to estimate the covariance matrix of $\hat\theta$. This procedure is valid only if the model is correctly specified. When some explanatory variables are omitted from the model, $\mathrm{Cov}[\partial\log\ f(y_i|\theta_1^*,0)/\partial\theta_1]$ does not equal $E(\partial\log\ f^*/\partial\theta_1)(\partial\log\ f^*/\partial\theta_1')$ because $E[\partial\log\ f(y_i|\theta_1^*,0)/\partial\theta_1]$ is not zero; only $E[\partial\log L(Y;\theta_1^*,0)/\partial\theta_1]$ is zero by the definition of θ_1^*. In this case (3.15) - (3.17) should be used to estimate $\mathrm{Cov}(\hat\theta_1^*)$.

To apply our information criterion (3.16) to estimate $-E_{\hat\theta_1^*}I[g;f]$ for a linear regression model f , let the true model g be (2.1) and let f be the model with $\beta_2 = 0$. Using the definition (3.5), we maximize $E_{\tilde{Y}}[\log L(\tilde{Y}|\beta_1,0,\sigma^2)]$ with respect to β_1 and σ^2, assuming $\tilde{Y} = \tilde{X}\beta + \tilde{u}$, with $\tilde{X} = X$. The results are β_1^* as given by (3.8) and

$$(3.18) \qquad \sigma^{*2} = n^{-1}\beta_2' X_{2.1}' X_{2.1}\beta_2 + \sigma^2$$

where σ^2 is the true residual variance of (2.1). Denoting the partial derivatives of $\log L(\tilde{Y}|\beta_1,0,\sigma^2)$ with respect to β_1 and σ^2 evaluated at β_1^* and σ^{*2} simply by $\partial \log L^*$, one easily derives

$$(3.19) \qquad \frac{\partial \log L^*}{\partial \beta_1} = \frac{1}{\sigma^{*2}} X_1'(\tilde{Y} - X_1\beta_1^*) = \frac{1}{\sigma^{*2}} X_1'\tilde{u}$$

$$(3.20) \qquad \frac{\partial \log L^*}{\partial \sigma^2} = -\frac{n}{2\sigma^{*2}} + \frac{1}{2\sigma^{*4}}(\tilde{Y} - X_1\beta_1^*)'(\tilde{Y} - X_1\beta_1^*)$$

$$= -\frac{n}{2\sigma^{*2}} + \frac{1}{2\sigma^{*4}}(\beta_2' X_{2.1}' X_{2.1}\beta_2 + \tilde{u}'\tilde{u} + 2\tilde{u}' X_{2.1}\beta_2)$$

from which one deduces

$$(3.21) \qquad J_{11}(\theta^*,\theta^o) = -E\begin{bmatrix} \dfrac{\partial^2 \log L^*}{\partial\beta_1\partial\beta_1'} & \dfrac{\partial^2 \log L^*}{\partial\beta_1\partial\sigma^2} \\ \dfrac{\partial^2 \log L^*}{\partial\sigma^2\partial\beta_1'} & \dfrac{\partial^2 \log L^*}{\partial(\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\sigma^{*2}} X_1' X_1 & 0 \\ 0 & \dfrac{n}{2\sigma^{*4}} \end{bmatrix} ;$$

$$(3.22) \qquad \mathrm{Cov}\begin{bmatrix} \dfrac{\partial \log L^*}{\partial\beta_1} \\ \dfrac{\partial \log L^*}{\partial\sigma^2} \end{bmatrix} = \begin{bmatrix} \dfrac{\sigma^2}{\sigma^{*4}} X_1' X_1 & 0 \\ 0 & \dfrac{n\sigma^2(2\sigma^{*2} - \sigma^2)}{2\sigma^{*8}} \end{bmatrix}$$

In deriving (3.22), we have made use of $E(\tilde{u}'\tilde{u})^2 = (2n+n^2)\sigma^4$ and $E\tilde{u}(\tilde{u}'\tilde{u}) = 0$ because the elements $\tilde{u}_i$ of $\tilde{u}$ are normal and independent. Substituting

(3.21) and (3.22) into the trace of (3.16) we find, for the approximate regression model,

$$(3.23) \qquad \mathrm{tr}\{\mathrm{Cov}\left[\frac{\partial \log L^*}{\partial \theta_1}\right]\left[J_{11}(\theta^*, \theta^o)\right]^{-1}\} = \frac{\sigma^2}{\sigma^{*2}}\left[k_1 + 2 - \frac{\sigma^2}{\sigma^{*2}}\right]$$

where σ^2 is the residual variance of the true regression, σ^{*2} is the residual variance of the best approximate regression, k_1 is the number of coefficients in β_1, and $k_1 + 1$ is the number of parameters. The result (3.23) is identical with the estimate of $-E_{\hat{\theta}_1^*} I[g;f]$ for the linear regression model by Sawa (1978, Theorem 3.2, p. 1280). Our criterion (3.16) is more general. The required trace in (3.16), which depends on the unknown parameters of the approximate and true models as illustrated by (3.23), can be estimated by using the maximum likelihood estimates of these parameters.

A by-product of the derivation of (3.16) is the covariance matrix given by (3.15) for the asymptotic distribution of the maximum likelihood estimator when the model is incorrectly specified. It can be estimated by using (3.17). The analysis of this section could be extended to the case where the successive observations $\tilde{y}_i$ $(i = 1,\ldots,n)$ are not independent as long as they are generated by the same mechanism which has generated the sample observations y_i $(i = 1,\ldots,n)$. But this topic will not be pursued in this paper.

To close this section, let us compare Akaike's information criterion with the criterion (2.9) for the choice between the two nested models in section 2. The former criterion favors the small model if the maximum value of its likelihood is larger than the maximum value of the likelihood for the large model minus k_2, i.e., if

$$(3.24) \quad -\frac{n}{2}\log(\hat{\sigma}^{*2}/\hat{\sigma}^2) > -k_2$$

where

$$(3.25) \quad \hat{\sigma}^{*2} = \frac{1}{n}(Y-X_1\hat{\beta}_1^*)'(Y-X_1\hat{\beta}_1^*) = \hat{\sigma}^2 + \frac{1}{n}\hat{\beta}_2'X_{2.1}'X_{2.1}\hat{\beta}_2 .$$

If we approximate $\log(\hat{\sigma}^{*2}/\hat{\sigma}^2)$ roughly by $(\hat{\sigma}^{*2}/\hat{\sigma}^2)-1$ and use (3.25), (3.24) becomes

$$(3.26) \quad \hat{\beta}_2'X_{2.1}'X_{2.1}\hat{\beta}_2 < 2k_2\hat{\sigma}^2$$

which is nearly the same as the condition (2.9) derived from the expected squared prediction error criterion. One reason why Akaike's Information Criterion (with the rough approximation of $\log(\hat{\sigma}^{*2}/\hat{\sigma}^2)$ above) is so close to (2.9) in spite of its error in estimating $\text{Cov}(\hat{\theta}_1^*)$ is that the expected squared prediction error criterion is itself only a rough approximation of expected information. The expected information criterion penalizes a regression model when it has an incorrect value for the residual variance whereas the expected squared prediction error criterion does not, as we have seen from (2.14).

4. Selection of Systems of Simultaneous Equations

In this section, we consider three problems arising from the selection among models of simultaneous equations. First, consider the choice between a simultaneous model

$$(4.1)\quad \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} + \begin{bmatrix} \Gamma_1 \\ \Gamma_2 \end{bmatrix} x_t = \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

and a block-recursive model obtained by the restrictions $B_{12} = 0$ and $\Sigma_{12} = Eu_{1t}u'_{2t} = 0$. The information criterion (3.16)-(3.17) can be applied if the method of (full-information) maximum likelihood is used to estimate the parameters of both models. A statistical criterion is thus provided to decide whether a system of simultaneous econometric equations should be decomposed into two recursive blocks. The issue is equivalent to deciding whether a general equilibrium or a partial equilibrium model should be selected to predict a subset y_{2t} of endogenous variables, the latter model treating y_{1t} as exogenous. The criterion is applicable even if the simultaneous model is nonlinear, as long as the recursive model is obtained by restricting a subset of its parameters to be zero.

The second problem is to choose between two simultaneous-equation models having different sets of endogenous variables. Let the first model be (4.1) and the second be a linear model having y_{1t} and y_{3t} as endogenous variables. To make this problem meaningful, we propose that the criterion be better prediction of a common subset y_{1t} of endogenous variables. For example, one may ask which model can better predict GNP, the inflation rate and the unemployment rate as measured by expected information. The two models may differ a great deal in size, one being very aggregative, for instance. To solve this problem, we consider the reduced-form of each model. Under the assumption of normal, serially uncorrelated residuals, the log-likelihood function of y_{1t} via its reduced form (for Model A say, with subscript A omitted) is

$$(4.2)\quad \log L(Y_1;\theta) = \text{const} - \frac{n}{2}\log|\Omega_{11}| - \frac{1}{2}\sum_{t=1}^{n}(y_{1t}-\Pi_1x_t)'\Omega_{11}^{-1}(y_{1t}-\Pi_1x_t)$$
$$= \text{const} - \frac{n}{2}\log\left|\frac{1}{n}\sum_{t=1}^{n}(y_{1t}-\Pi_1x_t)(y_{1t}-\Pi_1x_t)'\right|$$

where the second line gives the concentrated log-likelihood after the covariance matrix Ω_{11} is eliminated by step-wise maximization. If the reduced-form coefficients Π_1 are estimated by maximizing (4.2) without imposing the possible over-identifying restrictions due to the specification of the structure, the information criterion (3.16)-(3.17) can be applied to the selection between two models A and B.

The third problem is a version of the second problem when the parameters Π_1 in the reduced-form for y_{1t} are estimated by maximum likelihood subject to the over-identifying restrictions. The derivation of the information criterion (3.16)-(3.17) in section 3 has to be modified to account for these restrictions. To retrace the four steps in section 3, let the parameter vector θ of the true model include the coefficients Π_1 of the exogenous variables in Model A , the coefficients Π_2 of the exogenous variables in Model B but not Model A, and a covariance matrix Ω of the residuals. Model A is thus an approximation of a true multivariate regression model by restricting a subset of its parameters to be zero. Its parameter vector θ_1 consists of the elements of Π_1 and those of a covariance matrix Ω_{11} . In addition, however, θ_1 is subject to a set of restrictions $h(\theta_1) = 0$ which are due to the overidentified structural equations of Model A. (Strictly speaking, the arguments of $h(\cdot)$ include also the coefficients of the reduced-form equations explaining the other endogenous variables in Model A than y_{1t} , but we treat these arguments as constants equal to their maximum likelihood estimates.)

Define θ^o as the parameters of the true reduced-form, θ_1^* as the (pseudo-true) parameters of the best approximate Model A, and $\hat{\theta}_1^*$ as the maximum likelihood estimate of θ_1^* subject to the restrictions $h(\theta_1) = 0$. The quadratic approximation (3.3) still applies to $I[g;f(\cdot|\hat{\theta}_1^*,0)]$, but

(3.7) no longer holds for θ_1^* because (3.6) will have to be minimized subject to the restrictions $h(\theta_1) = 0$. The second line of (3.9) will become, with θ^* denoting $(\theta_1^*,0)$,

$$(4.3)\quad I[g;f(\cdot|\hat{\theta}_1^*,0)] = \frac{1}{2}(\hat{\theta}_1^*-\theta_1^*)'J_{11}(\hat{\theta}_1^*-\theta_1^*)+\frac{1}{2}(\theta^*-\theta^o)'J(\theta^*-\theta^o)+(\theta^*-\theta^o)'J\begin{bmatrix}\hat{\theta}_1^*-\theta_1^*\\ 0\end{bmatrix}$$

where the last cross-product term remains because (3.7) no longer holds. In the fourth step, we try to estimate the second term $\frac{1}{2}(\theta^*-\theta^o)'J(\theta^*-\theta^o)$ by (3.10) as before. Equation (3.11) will include one extra term because $\partial\log L(Y;\hat{\theta}_1^*,0)/\partial\theta_1$ no longer vanishes; it becomes

$$(4.4)\quad E_{\tilde{Y}}\log L(\tilde{Y}|\theta^*) \simeq \log L(Y;\hat{\theta}_1^*,0) - \frac{1}{2}(\hat{\theta}_1^*-\theta_1^*)'J_{11}(\hat{\theta}_1^*-\theta_1^*) + \frac{\partial\log L(Y;\hat{\theta}_1^*,0)}{\partial\theta_1}(\hat{\theta}_1^*-\theta_1^*) .$$

Combining (4.3), (3.10) and (4.4), we obtain an equation similar to (3.12), except that there are two extra terms involving $(\hat{\theta}_1^*-\theta_1^*)$ which are respectively derived from the last terms of (4.3) and (4.4). However, when we take expectations of this equation to estimate $-E_{\hat{\theta}_1^*}I[g;f(\cdot|\hat{\theta}_1^*,0)]$, these extra terms can be ignored because $\hat{\theta}_1^*$ subject to the restrictions $h(\hat{\theta}_1^*) = 0$ is an asymptotically unbiased estimator of θ_1^* according to Lemma 1 of Silvey (1959, p. 394). Therefore, the model selection criterion (3.13) remains valid and can be rewritten as

$$(4.5)\quad \log L(Y;\hat{\theta}_1^*,0) - \mathrm{tr}\{ J_{11}(\theta^*,\theta^o)\cdot\mathrm{Cov}(\hat{\theta}_1^*)\} - E\log L(\tilde{Y}|\theta^o) .$$

The covariance matrix of the asymptotic distribution of $\hat{\theta}_1^*$ is given by Lemma 1 of Silvey (1959). (The subject of constrained maximum likelihood estimation is further studied by Rothenberg (1973) and Wegge (1978) in the context of simultaneous equations.) We can continue to estimate $J_{11}(\theta^*,\theta^o)$ by the expression on the left side of (3.17). Thus, in theory the problem of estimating the mean information of an approximate linear reduced-form model for predicting a subset y_{1t} of endogenous variables can be solved by using (4.5), where the estimate $\hat{\theta}_1^*$ is obtained by the method of maximum likelihood subject to the overidentifying restrictions. We have not dealt with the computational problems involved in estimating $\mathrm{Cov}(\hat{\theta}_1^*)$ by Silvey's formula and in estimating $J_{11}(\theta^*,\theta^o)$ by (3.17). Nor have we examined the estimation errors which may result from using the covariance matrix of only the asymptotic distribution of $\hat{\theta}_1^*$ and from replacing the parameters in this covariance matrix by the maximum likelihood estimates.

5. Comparison with the Posterior Probability Criterion

Having explained the expected information criterion and discussed some of its applications, we would like to compare briefly its logic with that of the posterior probability criterion. The latter has been advocated by Jeffreys (1961) and adopted by Zellner (1971), S. Geisel (1975), Schwarz (1978) and Leamer (1978), among others. In choosing from models $M_j (j=1,...,J)$ and assuming a symmetric loss function, one selects the model with the highest posterior probability $p(M_j|Y)$ given the data $Y=(y_1,...,y_n)$. By the Bayes theorem ,

$$(5.1)\quad p(M_j|Y) = \frac{p(M_j,Y)}{p(Y)} = \frac{p(Y|M_j)p(M_j)}{\sum_j p(Y|M_j)p(M_j)}$$

Since the denominator $p(Y)$ is common to all models, if the prior probability $p(M_j)$ is the same for all models, the posterior probability $p(M_j|Y)$ is proportional to the likelihood $p(Y|M_j)$ of the data Y given the model M_j , which is evaluated by

$$(5.2)\quad p(Y|M_j) = \int L_j(Y;\theta)p_j(\theta)d\theta$$

where $L_j = \prod_{i=1}^{n} f_j(y_i|\theta)$ is the likelihood function of the model and $p_j(\theta)$ is the prior density of the parameter vector of the j^{th} model.

Schwarz (1978) has tried to estimate (5.2) for large samples and obtained a different formula for model selection from Akaike's formula based on the information criterion. The difference occurs because the definitions of the word "models" used by these two selection criteria are different. When the posterior probability criterion is applied, the model M_j refers to the density function $f_j(y|\theta)$ together with the prior density $p_j(\theta)$, as it can be seen from (5.2). The criterion is used to decide whether the model so defined is in agreement with the sample $Y=(y_1,\dots,y_n)$. This is done by evaluating the likelihood $p(Y|M_j)$ of the sample Y given the model M_j using (5.2). The sample Y is used to judge the model which was specified before the sample is available.

By contrast, as the expositions of sections 2 and 3 have made clear, when the information criterion is applied, the model M_j refers to the density function $f_j(y|\hat{\theta})$ where the parameter value $\hat{\theta}$ is estimated by the sample data Y . It is a post-sample model, whereas the posterior probability criterion is concerned with a pre-sample model. The sample Y is now used to estimate $\hat{\theta}$ in the model, and not to judge whether a pre-sample model is good. The post-sample model $f_j(\cdot|\hat{\theta})$ will be judged by predictions of future observations not yet available. The logic of this post-sample prediction criterion was made clear in section 2 where the criterion was the accuracy of predicting the observations $\tilde{Y}$ of (2.2) which are not yet available. The only change in section 3 was to use the information concept rather than the mean squared prediction error to measure the accuracy of prediction.

In summary, if the data are used to estimate the model and the estimated model $f_j(\cdot|\hat{\theta})$ is to be judged by new observations, the information criterion is relevant. Surely, the information concept could be applied to measure how well a pre-sample model fits the sample data Y , but the information criterion as adopted by Akaike (1973, 1974) and described in this paper refers to its application to measure future predictions by a model $f_j(\cdot|\hat{\theta})$, $\hat{\theta}$ having been estimated by the sample data Y . On the other hand, if a model is defined by the density function $f_j(\cdot|\theta)$ where θ is specified by some prior density $p_j(\theta)$ without using the sample Y , and if the sample Y is used to judge this pre-sample model, the posterior-probability criterion is relevant.

Let M_2 be a large model and M_1 be a smaller model obtained by restricting some parameters of M_2 to be zero. Let the pre-sample information consist of 50 observations which were used to obtain the prior densities $p_1(\theta_1)$ and $p_2(\theta)$ of the parameters of these two models. The current sample Y consists of 100 observations. The pre-sample M_1 might be a better-predictive model than the pre-sample M_2 because it is smaller; the additional parameters in M_2 caused larger sampling errors in the parameter estimates and could generate larger errors in prediction as we have observed in the discussion following (2.9). If these two pre-sample models are judged by the sample Y , the conclusion obtained by the posterior-probability criterion might favor the small pre-sample model M_1 . However, once the sample data Y are used to estimate both models (with 150 observations being used in total), the larger post-sample model M_2 might be judged by the information criterion to be likely to yield better predictions in the future.

The usefulness of the information criterion lies in its ability to discriminate between models having different numbers of parameters. Its main weakness is that only a point estimate is used to estimate expected information (or expected predictive ability) without regard to its sampling errors. This weakness is apparent from the decision rule (2.11) for choosing between two non-nested regression models. The issue of sampling errors in the estimate deserves further investigation. When the numbers of parameters are equal in two models, the distinction between a pre-sample and a post-sample model becomes less important, as the discussion of the last paragraph suggests. Nevertheless, insofar as the current sample Y is already available, it may be desirable to use it to estimate the parameters of both models and compare (the point estimates of) their expected predictive ability using the information criterion. Note that the information criterion, like a test of significance, becomes invalid if the extra explanatory variables in a larger model are included after much data mining, as it is clear from our discussion of the selection criterion (2.9). Note also that the posterior probability criterion can be supplemented by using the prior probabilities $p(M_j)$ of the models and by considering an explicit loss-function. Hopefully these remarks are helpful to the reader in judging the applicability of these two criteria to the model selection problem at hand.

6. Concluding Remarks

In this paper, we have provided an exposition of the information criterion, suggested an estimate of expected information, illustrated its applicability to problems of econometric model selection, and compared its logic with that of the posterior probability criterion. Although the information criterion is far from being the final answer to the general problem of model selection, it is sufficiently promising to deserve the consideration of econometricians especially when the problem is to select from models having different numbers of parameters. It should not be surprising to find Bayesians who would argue in favor of the posterior-probability criterion more strongly than the author did in section 5. Furthermore, critics of the information criterion could point out that the pre-test procedures in sections 2 and 3 are inadmissible, as Sawa (1978, p. 1274) has recognized. Insofar as the procedures described are non-Bayesian, they might not appeal to the Bayesian statisticians. Nevertheless, this author believes that these procedures are of practical value and should be presented clearly to applied econometricians for their final judgment.

Our discussion has also generated several interesting problems for further study, including the estimation of the covariance matrix of a maximum likelihood estimator (3.15) when the model has omitted some explanatory variables, the generalization of Section 3 to the cases when the successive observations are not independent and when the restrictions imposed by the approximate model f are other than the omission of variables, the problem of sampling errors of our point estimate of expected information, and the estimation of expected information for the selection of simultaneous-equation models which are estimated by methods other than maximum likelihood. As the last remark, although the method of Section 3 is non-Bayesian, it could be supplemented by a Bayesian analysis. Once the measure of expected information is converted into a function of the parameters θ^o of the true model as obtained by taking the expectation of (3.9) and using (3.7), Bayesian parameter estimation theory could be applied to its estimation, which is another topic for further research.

REFERENCES

[1] Anderson, T. W.: Introduction to Multivariate Statistical Analysis. New York: John Wiley & Sons, 1958.

[2] Akaike, H.: "Information Theory and Extension of the Maximum Likelihood Principle," in Proc. 2nd Int. Symp. Information Theory, ed. by B. N. Petrov and F. Csáki. Budapest: Akademiai Kiadó, 267-281.

[3] __________: "A New Look at the Statistical Model Identification," IEEE Transactions on Automatic Control, AC-19 (1974), 716-723.

[4] Berndt, E. K., B. H. Hall, R. E. Hall and J. A. Hausman: "Estimation and Inference in Nonlinear Structural Models," Annals of Economic and Social Measurement, 3 (1974), 653-666.

[5] Chow, G. C.: "A Reconciliation of the Information and Posterior Probability Criteria for Model Selection," Econometric Research Program, Research Memorandum No. 234, (1979), Princeton University.

[6] Geisel, M. S.: "Bayesian Comparisons of Simple Macroeconomic Models," in Studies in Bayesian Econometrics and Statistics, ed. by S. E. Feinberg and A. Zellner, 1975. Amsterdam: North-Holland, 227-256.

[7] Jeffreys, H.: Theory of Probability, (3rd ed.). Oxford: Clarendon, 1961.

[8] Kullback, S.: Information Theory and Statistics. New York: John Wiley & Sons, 1959.

[9] Kullback, S. and R. A. Leibler: "On Information and Sufficiency," Annals of Mathematical Statistics, 22 (1951), 79-86.

[10] Leamer, E.: Specification Searches. New York: John Wiley & Sons, 1978.

[11] Mallows, C. L.: "Some Comments on C_p," Technometrics, 15 (1973), 661-675.

[12] Rao, C. R.: Linear Statistical Inference and its Applications, 2nd ed., New York: John Wiley & Sons, 1973.

[13] Rothenberg, T. J.: Efficient Estimation with A Priori Information, Cowles Foundation Monograph 23. New Haven: Yale University Press, 1973.

[14] Sawa, T.: "Information Criteria for Discriminating Among Alternative Regression Models," Econometrica, 46 (1978), 1273-1292.

[15] Savage, L. J.: The Foundations of Statistics. New York: John Wiley & Sons, 1954.

[16] Schwarz, G.: "Estimating the Dimension of a Model," Annals of Statistics, 6 (1978), 461-464.

[17] Silvey, S. D.: "The Lagrangian Multiplier Test," Annals of Math. Statistics, 30 (1959), 389-407.

[18] Wegge: "Constrained Indirect Least Squares Estimators," Econometrica, 46 (1978), 435-449.

[19] Zellner, A.: An Introduction to Bayesian Inference in Econometrics. New York: John Wiley & Sons, 1971.