A COMPARISON OF THE INFORMATION AND POSTERIOR PROBABILITY CRITERIA FOR MODEL SELECTION

Gregory C. Chow*

Econometric Research Program
Research Memorandum No. 253

November 1979

Abstract: This paper proposes an improvement of Akaike's information criterion for model selection. It provides a simple proof of Schwarz' formula for the posterior probability of a model being correct and points out its approximation error. It compares the two criteria from the viewpoint of statistical decision theory.

* I would like to acknowledge financial support from the National Science Foundation through Grant No. SOC77-07677.

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

A COMPARISON OF THE INFORMATION AND POSTERIOR PROBABILITY CRITERIA FOR MODEL SELECTION

By Gregory C. Chow
Department of Economics, Princeton University

Summary

This paper proposes an improvement of Akaike's information criterion for model selection. It provides a simple proof of Schwarz' formula for the posterior probability of a model being correct and points out its approximation error. It compares the two criteria from the viewpoint of statistical decision theory.

Some key words: Bayesian procedure; Information criterion; Model selection.

1. Introduction

The purposes of this paper are three-fold. First, it proposes an improvement of the formula of H. Akaike (1973, 1974) for model selection based on the information criterion. Second, it presents a simpler and more transparent derivation of the formula of G. Schwarz (1978) for model selection based on the posterior probability criterion, and points out its approximation error. Third, it compares these two model selection criteria from the viewpoint of statistical decision theory. These topics will be discussed respectively in sections 2, 3 and 4.

While the first two topics are technical in nature and hopefully

noncontroversial, the third topic is somewhat philosophical and possibly controversial as it touches upon the foundation of statistical inference. However, such a discussion is unavoidable if a practitioner is to decide intelligently which of the two conflicting formulas should be used.

The question to be studied is the following. Given J models represented by the densities $f_1(\cdot|\theta_1),\ldots,f_J(\cdot|\theta_J)$ for the explanation of a random vector y, and given n observations, how should one model be selected as being the best? To make our study manageable, we make two further assumptions. First, there exists a general model $f(\cdot|\theta)$ from which all the J competing models can be derived by imposing various restrictions on its parameter vector θ. Second, such a model is the "true" model generating the observations $y_1,\ldots,y_n$. We will let it be the first model for convenience. It is recognized that these assumptions are restrictive, as they rule out some important model-selection problems to which the posterior probability criterion has been applied. However, they do encompass the important classical statistical problems of testing the null-hypothesis that the parameter vector θ is subject to a set of restrictions and of choosing among several non-nested models provided that they can all be derived from restricting the parameters of a more general model.

2. Derivation of An Information Criterion

Let $f(\cdot|\theta^o)$ be the true density of y and $f(\cdot|\theta)$ be an approximation of $f(\cdot|\theta^o)$ where θ is subject to certain restrictions which θ^o does not satisfy. Following Akaike (1973, 1974), we adopt the Kullback-Leibler information measure, or the expected log-likelihood

ratio, to discriminate between the two models using n future independent observations $(\tilde{y}_1,\ldots,\tilde{y}_n) = \tilde{Y}$:

$$I_n[\theta^o;\theta] = E \sum_{i=1}^{n} [\log f(\tilde{y}_i|\theta^o) - \log f(\tilde{y}_i|\theta)] \tag{2.1}$$

where the expectation is evaluated by the true density $f(\cdot|\theta^o)$. As θ is unknown, we assume that n observations $(y_1,\ldots,y_n) = Y$ are available to provide a maximum likelihood estimate $\hat{\theta}$ of θ subject to the required restrictions. The estimated model $f(\cdot|\hat{\theta})$ is to be judged by the expected information

$$E_{\hat{\theta}} I_n[\theta^o;\hat{\theta}] = E_{\hat{\theta}}[E_{\tilde{Y}} \log L(\tilde{Y};\theta^o) - E_{Y} \log L(\tilde{Y};\hat{\theta})] \tag{2.2}$$

where L denotes the likelihood function based on n future observations. Akaike (1973, 1974) has provided an estimate of $E_{\hat{\theta}}[E_{\tilde{Y}} \log L(\tilde{Y};\hat{\theta})]$ for model selection, since the term $E_{\tilde{Y}} \log L(\tilde{Y};\theta^o)$, though unknown, is the same for all approximate models. This section proposes an alternative estimate, under the assumption that θ is subject to the known linear restriction $H'\theta + b = 0$. The reader will recognize that generalization to the case of nonlinear restrictions on θ is straight-forward because of the work of Silvey (1959). Our derivation consists of five steps.

First, we approximate $I_n[\theta^o;\theta]$ by a quadratic form in $\theta-\theta^o$. Expanding $\log L(\tilde{Y};\theta)$ in a second-order Taylor series about θ^o and substituting the result into (2.1) we obtain, using the well-known fact $E[\partial \log L(\tilde{Y};\theta^o)/\partial\theta] = 0$,

$$I_n[\theta^o;\theta] = \tfrac{1}{2}(\theta-\theta^o)' J(\theta^+,\theta^o)(\theta-\theta^o) \tag{2.3}$$

where $\theta \leq \theta^{+} \leq \theta^{o}$ and

$$J(\theta^{+},\theta^{o}) \equiv - E_{\underset{Y}{\sim}} \frac{\partial^2 \log L(\tilde{Y};\theta^{+})}{\partial\theta\partial\theta'} \quad ,$$

the parameter θ^{o} of $J(\theta^{+},\theta^{o})$ being used to define the distribution of $\tilde{Y}$. $J(\theta^{o};\theta^{o})$ is Fisher's information matrix.

Second, given the linear restrictions $H'\theta + b = 0$, we find the best approximate model by minimizing the information $n^{-1}I_n[\theta^{o};\theta]$ with respect to θ subject to $H'\theta + b = 0$. Using (2.3) for I_n , we differentiate the Lagrangian expression, suppressing the arguments of J ,

$$\frac{1}{2n}(\theta-\theta^{o})'J(\theta-\theta^{o}) - \lambda'(H'\theta+b)$$

to yield

$$\begin{bmatrix} n^{-1}J & -H \\ -H' & 0 \end{bmatrix} \begin{bmatrix} \theta^{*} \\ \lambda^{*} \end{bmatrix} = \begin{bmatrix} n^{-1}J\theta^{o} \\ b \end{bmatrix}$$

the solution of which is

$$\theta^{*} = \theta^{o} + nJ^{-1}H\lambda^{*} \; ; \; \lambda^{*} = - n^{-1}(H'J^{-1}H)^{-1}(H'\theta^{o} + b) \qquad (2.4)$$

The vector θ^{*} can be considered the parameter of the approximate model and is called the pseudo-true parameter of the pseudo-true model in the language of Sawa (1978).

Third, if $\hat{\theta}^{*}$ is the maximum likelihood estimate of the parameter θ^{*} of the approximate model, we substitute $(\hat{\theta}^{*}-\theta^{*}) + (\theta^{*}-\theta^{o})$ for $(\theta-\theta^{o})$ in (2.3) to obtain the information measure for the estimated model

$$I_n[\theta^o;\hat{\theta}^*] = \frac{1}{2}(\hat{\theta}^*-\theta^*)'J(\hat{\theta}^*-\theta^*) + \frac{1}{2}(\theta^*-\theta^o)'J(\theta^*-\theta^o) \qquad (2.5)$$

where the cross-product $(\hat{\theta}^*-\theta^*)'J(\theta^*-\theta^o)$ has vanished on account of (2.4) and $H'\theta^* = H'\hat{\theta}^* = -b$. Parenthetically, the method of maximum likelihood is justified as it chooses θ to minimize the information measure based on the sample $Y = (y_1,\ldots,y_n)$, i.e.,

$$\sum_{i=1}^{n} \log f(y_i|\theta^o) - \sum_{i=1}^{n} \log f(y_i|\theta) .$$

Fourth, given a sample $Y = (y_1,\ldots,y_n)$ of n observations, we will estimate the second term of (2.5) as follows.

$$\frac{1}{2}(\theta^*-\theta^o)'J(\theta^*-\theta^o) = I_n[\theta^o;\theta^*] = E\log L(\tilde{Y};\theta^o) - E\log L(\tilde{Y};\theta^*)$$

$$\simeq E\log L(\tilde{Y};\theta^o) - \log L(Y;\theta^*) \qquad (2.6)$$

where we have estimated $E\log L(\tilde{Y};\theta^*)$ by its sample analogue $\log L(Y;\theta^*)$. We will find the maximum likelihood estimate $\hat{\theta}^*$ and expand $\log L(Y;\theta^*)$ in a second-order Taylor series about $\hat{\theta}^*$. $\hat{\theta}^*$ is found by differentiating

$$n^{-1}\log L(Y;\theta) + \lambda'(H'\theta+b)$$

to yield

$$n^{-1}\frac{\partial \log L(Y;\hat{\theta}^*)}{\partial\theta} + H\hat{\lambda} = 0 .$$

Expanding $\log L(Y;\theta^*)$ in (2.6) about $\hat{\theta}^*$, we obtain

$$\frac{1}{2}(\theta^*-\theta^o)'J(\theta^*-\theta^o) \simeq E\log L(\tilde{Y};\theta^o) - \log L(Y;\hat{\theta}^*)$$

$$- \frac{1}{2}(\theta^*-\hat{\theta}^*)'\frac{\partial^2 \log L(Y;\theta^*)}{\partial\theta\partial\theta'}(\theta^*-\hat{\theta}^*)$$

where we have observed $(\theta^*-\hat{\theta}^*)'[\partial \log L(Y;\hat{\theta}^*)/\partial\theta] = 0$ on account of the likelihood equation for $\hat{\theta}^*$ and $(\theta^*-\hat{\theta}^*)'H = 0$.

If $-\partial^2 \log L(Y;\theta^*)/\partial\theta\partial\theta'$ is replaced by its expectation $J(\theta^*,\theta^o)$, and the above equation is combined with (2.5), the result is

$$I_n[\theta^o;\hat{\theta}^*] \simeq E\log L(\tilde{Y};\theta^o) - \log L(Y;\hat{\theta}^*) + (\hat{\theta}^*-\theta^*)'J(\theta^*,\theta^o)(\hat{\theta}^*-\theta^*).$$

Since $E\log L(\tilde{Y};\theta^o)$, though unknown, is constant among alternative models obtained by specifying different sets of restrictions on θ, it can be ignored for the purpose of model selection.

Fifth, we arrive at a criterion for model selection by taking the expectation of $-I_n[\theta^o;\hat{\theta}^*]$ over the sampling distribution of $\hat{\theta}^*$, (plus the above constant term), i.e.

$$- E_{\hat{\theta}^*} I_n[g;f(.|\hat{\theta}^*)] + E\log L(\tilde{Y};\theta^o)$$

$$\simeq \log L(Y;\hat{\theta}^*) - \mathrm{tr}\{J(\theta^*,\theta^o)E(\hat{\theta}^*-\theta^*)(\hat{\theta}^*-\theta^*)'\}. \qquad (2.7)$$

The models will be ranked by (2.7), the one having the highest value to be selected. The remaining problem is to provide estimates of $J(\theta^*,\theta^o)$ and $E(\hat{\theta}^*-\theta^*)(\hat{\theta}^*-\theta^*)$.

To find the distribution of the maximum likelihood estimator subject to the restrictions $H\theta^* = -b$, we use the result of Silvey

(1959, Lemma 1). The joint distribution of $(\hat{\theta}^*-\theta^*)$ and the Lagrangean multiplier $(\hat{\lambda}-\lambda^*)$ are asymptotically normal with mean 0 and covariance matrix equal to n^{-1} times

$$\begin{bmatrix} P_{\theta*} V_{\theta*} P_{\theta*} & P_{\theta*} V_{\theta*} Q_{\theta*} \\ Q'_{\theta*} V_{\theta*} P_{\theta*} & Q'_{\theta*} V_{\theta*} Q_{\theta*} \end{bmatrix}$$

where, denoting $L(Y;\theta^*)$ by L^*,

$$nV_{\theta*} = E\left[\frac{\partial \log L^*}{\partial \theta} \cdot \frac{\partial \log L^*}{\partial \theta'}\right] - \frac{\partial E\log L^*}{\partial \theta} \cdot \frac{\partial E\log L^*}{\partial \theta'}$$

and

$$\begin{bmatrix} P_{\theta*} & Q_{\theta*} \\ Q'_{\theta*} & R_{\theta*} \end{bmatrix} = \begin{bmatrix} n^{-1}J(\theta^*,\theta^o) & -H \\ -H' & 0 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} nJ^{-1} - nJ^{-1}H(H'J^{-1}H)^{-1}H'J^{-1} & -J^{-1}H(H'J^{-1}H)^{-1} \\ -(H'J^{-1}H)^{-1}H'J^{-1} & -n^{-1}(H'J^{-1}H)^{-1} \end{bmatrix}.$$

In the important special case when the restrictions consist entirely of zero restrictions on a subset of parameters, we write $\theta^* = (\theta_1^* \; 0)$, $H' = [0 \quad I]$, and

$$J(\theta^*,\theta^o) = \begin{bmatrix} J_{11}(\theta^*,\theta^o) & J_{12}(\theta^*,\theta^o) \\ J_{21}(\theta^*,\theta^o) & J_{22}(\theta^*,\theta^o) \end{bmatrix}$$

The matrix $P_{\theta *}$ above becomes

$$P_{\theta *} = \begin{bmatrix} nJ_{11}^{-1}(\theta^*,\theta^o) & 0 \\ 0 & 0 \end{bmatrix}$$

and the covariance matrix of $(\hat{\theta}_1^* - \theta_1^*)$ becomes

$$J_{11}^{-1}(\theta^*,\theta^o)\left[E\frac{\log L^*}{\partial\theta_1}\cdot\frac{\log L^*}{\partial\theta_1'}\right]J_{11}^{-1}(\theta^*,\theta^o)$$

since $\partial E\log L^*/\partial\theta_1 = 0$ as θ_1^* is obtained by maximizing (differentiating) $E\log L(Y;\theta_1,0)$ with respect to θ_1. Using this result for the covariance matrix of $\hat{\theta}^*$ in (2.7), we have the following model selection criterion in the case of zero restrictions:

$$\log L(Y;\hat{\theta}^*) - \mathrm{tr}\{E\left[\frac{\partial\log L^*}{\partial\theta_1}\cdot\frac{\partial\log L^*}{\partial\theta_1'}\right]J_{11}^{-1}(\theta^*,\theta^o)\} \qquad (2.8)$$

Akaike (1973) was incorrect in claiming that $J_{11}^{-1}(\theta^*,\theta^o)$ is the asymptotic covariance matrix of $\hat{\theta}_1^*$, as we have just shown. If this claim were valid, the trace term in (2.7) would become k, the number of unknown parameters in θ_1, and (2.7) would become Akaike's information criterion by which one selects the model having the largest value for the maximum log-likelihood minus the number of parameters to be estimated. The claim is incorrect because only when the model is correctly specified, i.e., when $\theta^* = \theta^o$, do we have $J_{11}^{-1}(\theta^*,\theta^*)$ as the asymptotic covariance matrix of $\hat{\theta}^*$.

To illustrate the error in approximating the trace of (2.7) by k, consider the example of a true normal linear regression model for n

observations

$$Y = X_1\beta_1^o + X_2\beta_2^o + u = X\beta^o + u \qquad (\text{Cov } u = I\sigma^{o2})$$

which is being approximated by the smaller model

$$Y = X_1\beta_1^* + u^*.$$

The pseudo-true parameters β_1^* and σ^{*2} can be obtained by maximizing $E_{\tilde{Y}}[\log L(\tilde{Y}|\beta_1,0,\sigma^2)]$ with respect to β_1 and σ^2, where the new observations to be predicted are assumed to satisfy $\tilde{Y} = X\beta^o + \tilde{u}$. The results are

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

$$\sigma^{*2} = n^{-1}\beta_2'X_2'[I-X_1(X_1'X_1)^{-1}X_1']X_2\beta_2 + \sigma^{o2}.$$

Denoting the partial derivatives of $\log L(Y|\beta_1,0,\sigma^2)$ with respect to β_1 and σ^2 evaluated at β_1^* and σ^{*2} by $\partial\log L^*$, one finds

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

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

$$= -\frac{n}{2\sigma^{*2}} + \frac{1}{2\sigma^{*4}}(\beta_2^{o\prime}X_2'M_1X_2\beta_2^o + \tilde{u}'\tilde{u} + 2\tilde{u}'M_1X_2\beta_2^o)$$

where M_1 denotes $I - X_1(X_1'X_1)^{-1}X_1'$. From the last two equations, one derives

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

$$E\begin{bmatrix} \frac{\partial \log L^*}{\partial\beta_1} \\ \frac{\partial \log L^*}{\partial\sigma^2} \end{bmatrix}\begin{bmatrix} \frac{\log L^*}{\partial\beta_1'} & \frac{\log L^*}{\partial\sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{\sigma^{o2}}{\sigma^{*4}} X_1'X_1 & 0 \\ 0 & \frac{n\sigma^{o2}(2\sigma^{*2}-\sigma^{o2})}{2\sigma^{*8}} \end{bmatrix}.$$

The derivation of the last equation has made use of the relations $E(\tilde{u}'\tilde{u})^2 = (2n+n^2)\sigma^{o4}$ and $E\tilde{u}(\tilde{u}'\tilde{u}) = 0$ because the elements of $\tilde{u}_i$ of $\tilde{u}$ are normal and independent.

Using these results for the trace in (2.7), we get, for the approximate model

$$\mathrm{tr}\{E\left[\frac{\partial \log L^*}{\partial\theta_1} \cdot \frac{\partial \log L^*}{\partial\theta_1'}\right][J_{11}(\theta^*,\theta^o)]^{-1} = (\sigma^o/\sigma^*)^2[k+1-(\sigma^o/\sigma^*)^2]$$

where k is the number of parameters (including the elements of β_1^* and σ^{*2}). If the approximate model were true, $\sigma^* = \sigma^o$ and the trace term would equal k as Akaike claims. In general $\sigma^* > \sigma^o$, and the trace term is smaller than k. For example, if the true model contains eight parameters (seven coefficients plus σ^o) and the approximate model contains seven parameters (with the last explanatory variable omitted), and if $(\sigma^o/\sigma^*)^2 = .9$, the difference between the two trace terms to

be subtracted from the respective maximum log-likelihood functions is $8 - .9[8-.9] = 1.61$, as compared with $8-7 = 1$ by Akaike's formula. The selection rule (2.7) turns out to favor the small model more than Akaike's rule. The above adjustment constant is in agreement with the result of Sawa (1978, Theorem 3.2, p. 1280) who has studied the information criterion for the selection of linear regression models in particular. Our formula (2.7) has more general applicability. Note that to evaluate $E[\frac{\partial \log L^*}{\partial \theta_1} \cdot \frac{\partial \log L^*}{\partial \theta_1'}]$ and $J_{11}(\theta^*, \theta^o)$, one needs to specify the true model as the most general of the models to be selected and replace the required parameters θ^* and θ^o by their maximum likelihood estimates.

3. THE POSTERIOR PROBABILITY CRITERION

To state the Jeffreys-Bayes posterior probability criterion, let $p(M_j)$ be the prior probability for model M_j to be correct, and $p(\theta|M_j)$ be the prior density for the k_j-dimensional parameter vector θ_j conditioned on M_j being correct. Assume that a random sample of n observations $(y_1, y_2, \ldots, y_n) = Y$ is available. By Bayes' theorem the posterior probability of the j-th model being correct is

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

where

$$p(Y|M_j) = \int L_j(Y,\theta)p(\theta|M_j)d\theta \qquad (3.1)$$

with $L_j(Y,\theta_j)$ denoting the likelihood function for the j-th model.

Since $p(Y)$ is a common factor for all models, the model with the highest posterior probability of being correct is the one with the maximum value for

$$p(M_j)p(Y|M_j) = p(M_j)\int L_j(Y,\theta)p(\theta|M_j)d\theta .$$

If the prior probabilities $p(M_j)$ are equal for the models, the one with the highest $p(Y|M_j)$ will be selected.

To evaluate $p(Y|M_j)$ for large samples, we apply a well-known theorem of Jeffreys (1961, p. 193 ff), cited in Zellner (1977, pp. 31-33), on the posterior density $p(\theta|Y,M_j)$ of θ_j given model M_j:

$$p(\theta|Y,M_j) = \frac{L_j(Y,\theta)p(\theta|M_j)}{p(Y|M_j)} =$$

$$= (2\pi)^{-\frac{k_j}{2}} |S|^{\frac{1}{2}} e^{-\frac{1}{2}(\theta-\hat{\theta}_j)'S(\theta-\hat{\theta}_j)} [1+O(n^{-\frac{1}{2}})]$$

where $\hat{\theta}_j$ is the maximum likelihood estimate of θ_j and the inverse covariance matrix $S = -(\partial^2 \log L_j/\partial\theta\partial\theta')_{\hat{\theta}} \equiv nR_j$ is of order n. Evaluating both sides of the above equation at $\theta = \hat{\theta}_j$ and taking natural logarithms, we obtain

$$\log p(Y|M_j) = \log L_j(Y,\hat{\theta}_j) - \frac{k_j}{2}\log n - \frac{1}{2}\log|R_j|$$

$$+ \frac{k_j}{2}\log(2\pi) + \log p(\hat{\theta}_j|M_j) + O(n^{-\frac{1}{2}}) \qquad (3.2)$$

If we retain only the first two terms $\log L_j(Y,\hat{\theta}_j)$ and $-k_j(\frac{1}{2}\log n)$ in (3.2), we obtain the formula of Schwarz (1978).

How well can $\log p(Y|M_j)$ be approximated by using only the first two terms of (3.2)? How much will it depend on the prior density $p_j(\theta|M_j)$ of the parameter vector chosen for each model M_j? Bayesian statisticians including Jeffreys (1961), Pratt (1975), and Leamer (1978), among others, have recognized the difficult problem of choosing a prior distribution $p_j(\theta|M_j)$ for the parameters of each model to be used to compute $p(Y|M_j)$. The difficulty of this problem can be seen from the equation

$$p(Y|M_j) = \frac{L_j(Y,\hat{\theta}_j)p(\hat{\theta}_j|M_j)}{p(\hat{\theta}_j|Y,M_j)}$$

$$\simeq L_j(Y,\hat{\theta}_j)p(\hat{\theta}_j|M_j)\cdot(2\pi)^{\frac{k_j}{2}}|nR_j|^{-\frac{1}{2}}.$$

Observe that, given $L_j(Y,\hat{\theta}_j)$ and $p(\hat{\theta}_j|Y,M_j)$, $p(Y|M_j)$ is proportional to $p(\hat{\theta}_j|M_j)$. Thus one can change $p(Y|M_j)$ by a multiplicative factor simply by changing $p(\hat{\theta}_j|M_j)$ by that factor. If one wishes to use a diffuse prior density $p(\theta|M_j)$, many such densities are reasonable but they can give very different results. To illustrate, let $p(\theta|M_j)$ in (3.2) be k_j-variate normal with mean $\hat{\theta}_j$ (just for illustration) and covariance matrix $(\varepsilon R_j)^{-1}$. Equation (3.2) will become

$$\log p(Y|M_j) = \log L_j(Y,\hat{\theta}_j) - \frac{1}{2}k_j \log(\frac{n}{\varepsilon}) + O(n^{-\frac{1}{2}}). \qquad (3.3)$$

The adjustment constant suggested by the formula of Schwarz (1978) will be changed from $-\frac{1}{2}k_j \log n$ to $-\frac{1}{2}k_j \log(\frac{n}{\varepsilon})$. There is no reason why ε might not change by a factor of two or three, making Schwarz' formula a poor approximation to $\log p(Y|M_j)$ for finite samples.

4. Comparative Evaluation of the Two Criteria

We will begin by evaluating the information criterion from the viewpoint of Bayesian estimation theory, as Leamer (1979) has done. Under the assumption stated in the Introduction, the true model is the most general model $f(\cdot|\theta^o)$ with unknown parameter θ^o. $I[\theta^o;\theta]$ in (2.1) specifies a loss function for the approximate model $f(\cdot|\theta)$. $E_{\hat{\theta}_i} I[\theta^o;\hat{\theta}_i] = R_i(\theta^o)$ in (2.2) is the risk function for the estimator $\hat{\theta}_i$ which is subject to the restrictions defining the i-th model. Since the risk $R_i(\theta^o)$ depends on the specification of the model (i.e., the estimator) and the unknown θ^o, one cannot select the model (estimator) with minimum risk without knowing θ^o. A Bayesian will specify a prior density for θ^o, take the expectations $E_{\theta^o} R_i(\theta^o)$ $(i=1,\ldots,J)$, and choose the model i with the smallest expected risk. Instead, the proposal of section 2 is to evaluate $R_i(\hat{\theta}^o)$ using the maximum-likelihood estimator $\hat{\theta}^o$ of θ^o. This procedure appears ad hoc from the viewpoint of Bayesian estimation theory. Furthermore, since all Bayesian estimators defined by different prior densities on θ^o form a complete class of admissible estimators, and the above ad hoc procedure is not a Bayesian estimator, it is inadmissible.

A defense of the information criterion against the criticism from Bayesian estimation theory can be made as follows. First, if the risk $R_i(\theta^o)$ is adopted for ranking the i-th model or estimator, using a maximum likelihood estimate $R_i(\hat{\theta}^o)$ of it at least has large-sample justification from the viewpoint of sampling theory. Second, a Bayesian is challenged to provide an alternative procedure for model selection

which, from the sampling theory viewpoint, will on the average select a better model as judged by (2.1) than the information criterion. We will consider three Bayesian procedures below.

The first procedure is based on Bayesian estimation theory. Given a prior density on θ^o and given the loss function $I[\theta^o,\hat{\theta}]$, one can find an estimator $\hat{\theta}$ to minimize expected loss $E_{\theta^o}[\theta^o,\hat{\theta}]$ where the expectation is evaluated by the posterior density of θ^o. There are two problems with this procedure. First, it might not perform well from the sampling viewpoint. Second, it will never recommend imposing zero restrictions on any parameters, or dropping any explanatory variables in a regression model, unless the prior distribution of θ^o assigns probability one to these restrictions in the first place. Thus this procedure always leads to selecting the largest model in the problem formulated in the Introduction.

To justify the dropping of variables in statistical practice, two other Bayesian procedures can be mentioned, as discussed in Dickey (1975), for example. One involves introducing a reward for simplicity by subtracting a constant from the loss-function $\ell(\theta^o,\hat{\theta})$ when $\hat{\theta}$ satisfies the restrictions of a small model. The second is a Bayesian procedure for hypothesis testing. Given two hypotheses or models M_1 and M_2, it is required to specify a prior probability $p(M_i)$ for each model to be correct, a prior density $p_i(\theta_i|M_i)$ of the parameter θ_i for each model M_i, and a utility function $U(d;M)$ where d can take only two values d_i (for the decision to choose M_i), $i = 1, 2$. If M_1 stands for the general model with parameter $\theta = \theta^o$ and M_2 is obtained by restrictions on θ, the utility function can be written as $U(d;\theta)$. The

model M_i will be selected if $E_\theta U(d_i,\theta)$ is larger where the expectation is evaluated by the posterior density of θ. Note that this utility function is different from the loss function used in Bayesian estimation theory where the argument $\hat{\theta}$ is a continuous variable indicating the parameter estimate. Here d_i is a discrete variable referring to the decision to choose M_i, for an unspecified purpose except that $U(d_1,\theta^o) > U(d_2,\theta^o)$ and $U(d_2,\theta^*) > U(d_1,\theta^*)$ where M_2 is the restrictive model with parameter θ^*. Under the assumptions of a symmetrical utility function, i.e.,

$$U(d_1,\theta^o) = U(d_2,\theta^*) \quad \text{and} \quad U(d_2,\theta^o) = U(d_1,\theta^*) ,$$

this selection procedure amounts to selecting the model with the higher posterior probability $p(M_i|Y)$ of being correct. It further reduces to the selection by $p(Y|M_i)$ when $p(M_1) = p(M_2)$, as discussed in section 3.

What can be said about the posterior probability criterion for model selection? When applied to the choice between two nested models M_1 and M_2, the assumption of a symmetrical utility function becomes unreasonable since $U(d_1,\theta^*)$ depends on how far θ^* is from θ^o, and $U(d_1,\theta^o)$ cannot reasonably be set equal to $U(d_2,\theta^*)$ for all values of θ^o and θ^*. More importantly, the model M_i selected for having a higher value for $E_\theta U(d_i,\theta)$, as evaluated by the posterior density of θ, is not meant to be the model which, when estimated by maximum likelihood using a finite sample, will on the average predict future observations well by the information measure (2.1). Similarly, neither is the model having a higher $p(Y|M_i)$ meant to be the one which we should

estimate for prediction purposes. In our analysis, we have already assumed the most general or the largest model to be the true one, and yet imposing restrictions might produce a better model for prediction, given a finite sample.

To put the last point differently, the information criterion maximizes

$$E_{\hat{\theta}} E_{\tilde{Y}} \log L(\tilde{Y};\hat{\theta})$$

with the expectation evaluated by the sampling distribution of $\hat{\theta}$, whereas the posterior probability criterion maximizes

$$\log E_{\theta} L(Y;\theta) = \log \int L(Y;\theta) p(\theta|M_j) d\theta$$

with the expectation evaluated by the prior density of θ. This comparison brings out the basic difference between the two criteria as they attempt to answer two different questions. One asks which "model" $f(\cdot|\hat{\theta})$ as it is estimated by the given data Y should be used to predict the future $\tilde{Y}$. The other asks which "model" as defined by $f(\cdot|\theta)$ and the prior density $p(\theta|M)$ is judged by the sample data Y to have the highest probability of being correct. This distinction is not explicitly recognized in the literature. Schwarz (1978), in presenting his estimate of the posterior probability of a model being correct for large samples, stated that he was proposing an alternative formula to Akaike's for solving the same problem. Akaike (1978) asserted that he and Schwarz were trying to solve the same problem, and attempted to derive a formula close to his formula by using the posterior probability criterion. This could be done, for example, by choosing the prior

density

$$p(\hat{\theta}_j|M_j) = (2\pi)^{-\frac{k_j}{2}} |ne^{-2}R_j|^{\frac{1}{2}}$$

in (3.2) to make the entire adjustment factor equal to $-k_j$ instead of $-k_j(\frac{1}{2}\log n)$, but there is no need to justify the information criterion in terms of the posterior probability criterion as they are designed to answer different questions.

The above comparison also brings out the difficulty in choosing a robust prior density function $p(\theta|M_j)$ for the model selection problem. The "model" to be judged by the sample data Y using the posterior probability criterion is precisely defined by this prior density together with the function $f(\cdot|\theta)$. Varying the prior density $p(\theta|M_j)$ will vary significantly the "model" to be judged. Therefore, it is difficult to avoid choosing a specific prior density for the model selection problem using the posterior probability criterion. One might be tempted to resolve this difficulty by using a part Y_1 of the sample $Y = (Y_1\ Y_2)$ to obtain a preliminary $p(\theta|Y_1,M_j)$ from a diffuse $p(\theta|M_j)$, and then using the remaining data Y_2 to judge the "model" now specified by $p(\theta|Y_1,M_j)$ together with the function $f(\cdot|\theta)$. This suggestion can certainly be carried out, but it will answer the question whether the "model" based on the data Y_1 was good as judged by the data Y_2, and not whether the original model with a diffuse prior was good as judged by Y_1 and Y_2. Nor will it answer the interesting question whether the model estimated by using all the data Y will be good in future predictions.

In conclusion, although the information criterion is subject to

criticism from Bayesian estimation theory, it can be justified by sampling theory as it applies maximum likelihood to estimate the risk function $R_i(\theta^o)$. There are three Bayesian answers to the model selection problem posed in our Introduction. First, from Bayesian estimation theory with continuous loss and prior density functions, the largest model will always be selected, implying that explanatory variables should never be dropped from regression analysis. The estimator should take full account of the loss and prior density functions and not be restricted to maximum likelihood estimation of either the large or the small model, as is often done in statistical practice. The second answer justifies the selection of a smaller model by introducing discontinuity in the loss function (extra utility for imposing restrictions) and the third by introducing discontinuities in the prior density function and in the decision variable. If one accepts prediction as the criterion for model building, any of these three answers will have to be evaluated by its ability to produce good predictions.

I would like to thank Richard Quandt and James Trussel for helpful comments and to acknowledge financial support from the National Science Foundation.

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. Proceedings of the 2nd International Symposium on Information Theory, Ed. B. N. Petrov and F. Csáki, pp. 267-281. Budapest: Akademiai Kiadó.

_____ (1974). A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control, AC-19, pp. 716-723.

_____ (1978). A Bayesian Analysis of the Minimum AIC Procedure. Annals of the Institute of Statistical Mathematics, Part A, 30, pp. 9-14.

Dicky, J. (1975). Bayesian Alternatives to the F-test and Least-Squares Estimate in the Normal Linear Model. In Studies in Bayesian Econometrics and Statistics, Ed. S. E. Feinberg and A. Zellner, pp. 515-554. Amsterdam: North-Holland.

Geisel, M. S. (1975). Bayesian Comparisons of Simple Macroeconomic Models. In Studies in Bayesian Econometrics and Statistics, Ed. S. E. Feinberg and A. Zellner, pp. 227-256. Amsterdam: North-Holland.

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

Kullback, S. (1959). Information Theory and Statistics. New York: John Wiley and Sons.

Kullback, S. and R. A. Leibler (1951). On Information and Sufficiency. Annals of Mathematical Statistics, 22, pp. 79-86.

Leamer, E. (1978). Specification Searches. New York: John Wiley and Sons.

_____ (1979). Information Criterion for Choice of Regression Models. Econometrica, 47, pp. 507-510.

Pratt, J. W. (1975). Comments. In Studies in Bayesian Econometrics and Statistics, Ed. S. E. Feinberg and A. Zellner, pp. 71-73. Amsterdam: North-Holland.

Savage, L. J. (1954). The Foundations of Statistics. New York: John Wiley and Sons.

Sawa, T. (1978). Information Criteria for Discriminating Among Alternative Regression Models. Econometrica, 46, pp. 1273-1292.

Schwarz, G. (1978). Estimating the Dimension of a Model. Annals of Statistics, 6, pp. 461-464.

Silvey,S.D. (1959). The Lagrangian Multiplier Test.Ann.Math. Stat. 30, 389-407.

Zellner, A. (1971). An Introduction to Bayesian Influence in Econometrics. New York: John Wiley and Sons.