NO.416

MEAN SQUARED ERRORS OF FORECAST FOR SELECTING NON-NESTED LINEAR MODELS AND COMPARISON WITH OTHER CRITERIA

by

Hiroki Tsurumi and Hajime Wago

December 1988

converges. Equation (11) shows that the pdf of the MSEF is bounded by two chi-square distributions that are adjusted by maximum and minimum characteristic roots, μ_1 and μ_m, respectively, and equalities hold if $\mu_m = \mu_1$.

We shall use the distribution of the quadratic form in normal in deriving a predictive density of the MSEF. The predictive density of the MSEF is given by

$$\mathrm{p}(y_* \mid y, X, X_*) = \int_{-\infty}^{\infty} \int_0^{\infty} \mathrm{p}(y_* \mid \beta, \sigma, X_*)\mathrm{p}(\beta, \sigma \mid y, X)\, \mathrm{d}\sigma\, \mathrm{d}\beta \qquad (13)$$

The posterior pdf of β and σ^2 in (13) is given by

$$\mathrm{p}(\beta, \sigma^2 \mid y, X) \propto \sigma^{-(n+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[\nu s^2 + \left(\beta - \hat{\beta}\right)' X'X \left(\beta - \hat{\beta}\right)\right]\right\} \qquad (14)$$

which is obtained by using a diffuse prior $\mathrm{p}(\beta, \sigma) \propto \sigma^{-1}$. νs^2 in (14) is given by $\nu s^2 = y'[I - X(X'X)^{-1}X']y$, and $\nu = n - k$. y_* is distributed as $N(X_*\beta, \sigma^2 I_m)$. Integrating β out of (13) we have

$$\begin{aligned} \mathrm{p}(y_* \mid y, X, X_*) \propto \int_0^{\infty} \sigma^{-(\nu+m+1)} \exp\left(-\frac{\nu s^2}{2\sigma^2}\right) \\ \cdot \exp\left[-\frac{1}{2\sigma^2}\left(y_* - X_*\hat{\beta}\right)' \mathrm{H} \left(y_* - X_*\hat{\beta}\right)\right] \mathrm{d}\sigma^2 \end{aligned} \qquad (15)$$

where $\mathrm{H}=[I + X_*(X'X)^{-1}X_*']^{-1}$. The right-hand side of (15) shows that given σ, $y_* - X_*\hat{\beta}$ is distributed as $N(0, \sigma^2\mathrm{H}^{-1})$. Let $w = y_* - X_*\hat{\beta}$, and $\mathrm{H}=R'R$, where R is a nonsingular matrix. Then $m \cdot z = w'w = \eta'(RR')^{-1}\eta$, with $\eta = Rw \sim N(0, \sigma^2 I_m)$. Since the nonzero characteristic roots of $\mathrm{H}^{-1} = (R'R)^{-1}$ are the same as those of $(RR')^{-1}$, we see that $m \cdot z = \sum_{i=1}^{m} \mu_i \epsilon_i^2$, with $\epsilon_i \sim N(0, \sigma^2)$, and μ_i is the i-th characteristic root of H^{-1}. Hence, given σ, $m \cdot z$ has the distribution of a quadratic form in normal variables that is given in equation (4). Thus

$$\mathrm{p}(z \mid s^2, m) \propto \int_0^{\infty} \sigma^{-(\nu+m+1)} \mathrm{p}(z \mid \sigma^2, m) \exp\left(-\frac{\nu s^2}{2\sigma^2}\right) \mathrm{d}\sigma \qquad (16)$$

and integrating σ out, we obtain

$$\begin{aligned} \mathrm{p}(z \mid s^2, \nu, m) \quad \propto \quad & \frac{z^{\frac{m}{2}-1}}{\left(\nu s^2 + m\, z/\mu_m\right)^{(m+\nu)/2}} \\ & \cdot \sum_{p=0}^{\infty} \Gamma\left(\frac{m+\nu}{2} + p\right) 2^p \mathrm{c}(m, p) \left(\frac{m\, z}{\nu s^2 + m\, z/\mu_m}\right)^p \end{aligned} \qquad (17)$$

Using equation (11) we can show that the predictive pdf of z is bounded by two F distributions:

$$c_2 \frac{m^{\frac{m}{2}} z^{\frac{m}{2}-1}}{(\nu s^2 + m z/\nu_m)^{(m+\nu)/2}} \leq$$

$$\mathrm{p}(z \mid s^2, \nu, m) \leq c_2 \frac{m^{\frac{m}{2}} z^{\frac{m}{2}-1}}{(\nu s^2 + m z/\nu_1)^{(m+\nu)/2}} \tag{18}$$

where $c_2 = c_1 \Gamma(\frac{1}{2}) \Gamma(\frac{m+\nu}{2}) 2^{(m+\nu)/2-1} / \Gamma(\frac{m}{2})$. The equalities hold if $\mu_m = \mu_1$. When $m = 1$, the predictive density of the square root of the MSEF is identical to the predictive density for one period ahead forecast that is given, for example, in Zellner (1971, pp.72-73).

In Theorem 1 we used the degenerate hypergeometric function. The distribution of quadratic forms is often given by the Laguerre polynomials. If we arrange the Laguerre expansions given in Johnson and Kotz (1970, pp.159-160) to fit more conveniently to our case, the distribution of $x = m \cdot \mathrm{MSEF}/\sigma^2$ is given by

$$\begin{aligned} \mathrm{f}(x) &= \frac{1}{\gamma} \mathrm{p}_m(x/\gamma, 1) \Gamma(\frac{m}{2}) \sum_{p=0}^{\infty} \frac{(-2)^{-p}}{p! \Gamma(p + \frac{m}{2})} \\ &\quad \cdot \sum_{j=p}^{\infty} \frac{j!}{(j-p)!} c_j \gamma^{-j} (x/\gamma)^p \end{aligned} \tag{19}$$

where $\mathrm{p}_m(x/\gamma, 1)$ is the chi-square distribution with m degrees of freedom; $c_0 = 1$, $c_r = (2r)^{-1} \sum_{j=0}^{r-1} s_{r-j} c_j$, for $r \geq 1$; $s_r = \sum_{j=1}^{m} (\gamma - \mu_j)^r$, and γ is a number such that $\gamma > \frac{1}{2} \max_i \mu_i$.

If we use equation (19), the predictive density becomes

$$\begin{aligned} \mathrm{f}(z \mid s^2, \nu, \mu, \gamma) &\propto \frac{z^{\frac{m}{2}-1}}{(\nu s^2 + m z/\gamma)^{(m+\nu)/2}} \\ &\quad \sum_{p=0}^{\infty} \frac{(-1)^p \Gamma(\frac{m+\nu}{2})}{p! \Gamma(p + \frac{m}{2})} \sum_{j=p}^{\infty} \frac{j!}{(j-p)!} c_j^{-j} m^p \gamma^{-p} \left(\frac{z}{\nu s^2 + m z/\gamma} \right)^p \end{aligned} \tag{20}$$

Equation (20) is conditioned on one additional parameter, γ, whereas equation (17) is free of such an additional parameter. Moreover, our experience

suggests that equation (17) computationally yields much faster convergence than equation (20) due to the recursive coefficient $c(m,p)$.

Equation (17) is a Bayesian predictive density of the MSEF given s^2. A sampling distribution of the MSEF is given in Theorem 2 below:

Theorem 2: Let z be the MSEF. Then $u = z/\mu_m s^2$ has the probability density function given by

$$\mathrm{f}(u \mid \mu_1, \cdots, \mu_m, \nu) = \text{const} \cdot \frac{u^{\frac{m}{2}-1}}{\left(1+\frac{m\,u}{\nu}\right)^{(m+\nu)/2}} \sum_{p=0}^{\infty} \Gamma(\frac{m+\nu}{2})\, 2^p c(m,p)\, \mu_m^p \left(\frac{m}{\nu}\right)^p \left(\frac{u}{1+m\,u/\nu}\right)^p \tag{21}$$

where

$$\text{const} = \frac{(m/\nu)^{\frac{m}{2}} \mu_m^{\frac{m}{2}}}{\sqrt{\pi}\prod_{i=1}^{m} \mu_i^{1/2} \Gamma(\frac{\nu}{2})}$$

and $\nu\, s^2 = y' \left[I - X(X'X)^{-1}X'\right] y, \quad \nu = n-k.$

Proof: Equation (4) in Theorem 1 is the probability density function of $x = m \cdot \mathrm{MSEF}/\sigma^2$. The probability density function of $z = \mathrm{MSEF}$ may be obtained by transforming $z = (\sigma^2/m)x$, and this is given by equation (10). On the other hand, the pdf of s^2 is given by

$$\mathrm{f}(s^2 \mid \sigma^2, \nu) = \frac{\nu\,\sigma^{-2}}{2^{\frac{1}{2}}\Gamma(\frac{\nu}{2})} \left(\frac{\nu\, s^2}{\sigma^2}\right)^{\frac{\nu}{2}-1} \exp\left(-\frac{\nu\, s^2}{2\sigma^2}\right)$$

and it is easy to show that z and s^2 are independent. Hence, the joint pdf of z and s^2 is

$$\mathrm{f}(z,\, s^2 \mid m, \sigma^2) = c_3 \nu^{\frac{\nu}{2}} z^{\frac{m}{2}-1} (s^2)^{\frac{\nu}{2}} \sigma^{-(m+\nu)} \exp\left\{-\frac{\nu\, s^2}{2\sigma^2}\left[1 + m\, z/\nu\, \mu_m s^2\right]\right\} \sum_{p=0}^{\infty} c(m,p) m^p z^p \sigma^{-2p}$$

where $c_3 = c_1 \left[2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})\right]^{-1}$. Changing the variables z and s^2 to $u = z/\mu_m s^2$ and $y = s^2$ and integrating out y, we obtain equation (21).

3. Comparison of Certain Model Selection Criteria

The Bayesian predictive density of the MSEF that is given in (17) may be used as a criterion for selecting linear models. For each model we may draw the predictive density, and choose the model that has the mass of its density closest to zero. Or, we may choose the model that yields the smallest posterior mean of the MSEF. The choice of a quadratic loss function leads to the posterior mean of the MSEF as the point estimate of the MSEF.

The Bayesian criteria that are suggested above belong to the class of model selection criteria that are based on measures of how well each model explains data. Akaike's (1974) information criterion (AIC) and Efron's (1984) confidence interval for the mean squared errors also belong to this class. The AIC may be given by

$$\text{AIC} = n \log \hat{\sigma}^2 + 2k \tag{22}$$

where $\hat{\sigma}^2$ is the maximum likelihood estimate of σ^2, and k is the number of regression coefficients.

Efron's confidence interval for the difference of the mean squared errors of two competing models may be interpreted as an inferential procedure for the C_p that is suggested by Mallows (1973).

Let two linear regression models be given by

$$\begin{aligned} \text{Model A: } y &= W\gamma_A + X_A\beta_A + u = Z_A\delta_A + u \\ \text{Model B: } y &= W\gamma_B + X_B\beta_B + u = Z_B\delta_B + u \end{aligned} \tag{23}$$

where y is an $(n \times 1)$ vector of observations on the dependent variables; W is an $(n \times k_c)$ matrix of observations on the regressors that are common to both models; X_i is an $(n \times d_i)$ matrix of observations on the d_i regressors that are distinct to model i, $(i = A, B)$; $Z_i = (W, X_i)$; β_i is a $(d_i \times 1)$ vector of regression coefficients that are associated with the d_i distinct regressors of model i, while γ_i is a $(k_c \times 1)$ vector of regression coefficients that are associated with the k_c regressors that are common to both models; $\delta_i = (\gamma_i', \beta_i')'$, and u is an $(n \times 1)$ vector of error terms. The unbiased estimator of the difference of the mean squared errors (MSE) of models A

and B, $\Delta = \text{MSE}_B - \text{MSE}_A$, is given (after projecting out the common regressors, W)

$$\hat{\Delta} = \left(\mid y_{B^0} \mid^2 - \mid y_{A^0} \mid^2 \right) + 2(d_B - d_A)\bar{\sigma}^2 \tag{24}$$

where $\mid y_{B^0} \mid^2 = \hat{\beta}_A' X_A^{*'} M_B^* X_A^* \hat{\beta}_A$, $\mid y_{A^0} \mid^2 = \hat{\beta}_B' X_B^{*'} M_A^* X_B^* \hat{\beta}_B$, $M_i^* = M_w - X_i^{*'}(X_i^{*'} X_i^*)^{-1} X_i^*$, $X_i^* = M_w X_i$, $, M_w = I - W(W'W)^{-1}W'$ $i = A, B$; $\hat{\beta}_A$ and $\hat{\beta}_B$ are the least squares estimates of β_A and β_B in $M_w y = X_A^* \beta_A + X_B^* \beta_B + M_w u$, $\bar{\sigma}^2 = y'(I - ZZ^+)y/n$, $Z = (Z_A, Z_B)$. Efron decomposes the difference of the two MSE's, Δ, and its estimate, $\hat{\Delta}$, in a symmetric coordinate system and proposes to compute confidence intervals for Δ in the symmetric coordinate system. The computation of confidence intervals is suggested either by parametric bootstrapping or by non-parametric bootstrapping. Ideally the confidence interval for the difference of the MSE's of the two models, Δ, should be used for a hypothesis test for a specific value of Δ. In many cases, however, one does not possess knowledge of Δ, and is obliged to test the null hypothesis $\Delta = 0$. We shall adopt a strategy to reject model B in favor of model A if the 90% confidence interval for $\Delta = \text{MSE}_B - \text{MSE}_A$ lies entirely to the right of 0. One should realize that this strategy may lead to a low power of test around $\Delta = 0$ since it is much harder for the data to reject the hypothesis $\Delta = 0$ than the hypothesis that one of the model is true. We shall come back to this point later.

The model selection criteria that are discussed above are intended to settle the question of which of the two models is *better* rather than which is *correct*. In contrast, approaches that are based on Cox's tests of separate familites [Cox (1962)] are intended to settle the question of which model is correct and they are based on the translation of non-nested models into hypothesis testing on parameters. Pesaran (1974, 1982) proposes the N-tests (NT's). Davidson and MacKinnon (1981) suggest the J- tests (JT's). Tests based on the encompassing principle (ET's) are given in Mizon and Richard (1986). The NT's are given by

$$\text{N}_0 = \frac{n}{2} \log \left(\frac{\hat{\sigma}_B^2}{\hat{\sigma}_{BA}^2} \right) \left(\frac{\hat{\sigma}_A^2}{\hat{\sigma}_{BA}^4} \hat{\delta}_A' Z_A' M_B M_A M_B Z_A \hat{\delta}_A \right)^{-\frac{1}{2}} \tag{25}$$

where $\hat{\sigma}_i^2 = y' M_i y/n$, $M_i = I - Z_i(Z_i'Z_i)^{-1}Z_i'y$, $\hat{\delta}_i = (Z_i'Z_i)^{-1}Z_i'y$ $i = A, B$, and $\hat{\sigma}_{BA}^2 = \hat{\sigma}_A^2 + \hat{\delta}_A' Z_A' M_B M_A M_B Z_A \hat{\delta}_A/n$. The N_0 test is computed using

model A in (23) as the null hypothesis. The N_1 test is computed using model B as the null hypothesis, and it is given by interchanging subscripts A and B in (25).

The J tests (JT's) by Davidson and MacKinnon are the t-tests on parameter λ in

$$y = Z_A b_A + \lambda(Z_B \tilde{\delta}_B) + u \tag{26}$$

where $b_A = (1-\lambda)\delta_A$, and $\tilde{\delta}_B = (Z_B' Z_B)^{-1} Z_B' y$, and in

$$y = \lambda(Z_A \tilde{\delta}_A) + Z_B b_B + u \tag{27}$$

where $b_B = (1-\lambda)\delta_B$, and $\tilde{\delta}_A = (Z_A' Z_A)^{-1} Z_A' y$.

Mizon and Richard (1986) derive tests that are based on the encompassing principle. The encompassing test (ET) that is based on the conditional model is given by, if we use model A as the null-hypothesis,

$$\mathrm{ET}_0 = \frac{y' M_A X_B (X_B' M_A X_B)^{-1} X_B' M_A y}{y' M_c y} \cdot \frac{n - k_c - d_A - d_B}{d_B} \tag{28}$$

where $M_c = I - C(C'C)^{-1}C'$, and $C = (W, X_A, X_B)$. The numerator of (28) is identical to the Wald encompassing test that is derived by Mizon and Richard (1986) assuming that σ^2 is known. ET_0 is distributed as a central F with d_B and $n - k_c - d_A - d_B$ degrees of feedom if model A is correct. By using model B as the null-hypothesis, one obtains another encompassing test, ET_1, which is given by interchanging subscripts A and B in (28). For $d_A = d_B = 1$, the encompassing test (28) is identical to the JT's.

As is obvious from equation (25), the NT's can only be defined if $Z_A' M_B M_A M_B Z_A \neq 0$. Sufficient conditions for making this quantity zero are $M_B Z_A = 0$ or $M_B Z_A = Z_A$. $M_B Z_A = 0$ occurs if the columns of Z_A are linear combinations of the columns of Z_B, amd $M_B Z_A = Z_A$ occurs if $Z_B' Z_A = 0$ (i.e. when all the explanatory variables of the two models are orthogonal.) As for the JT's and ET's, they cannot be defined if linear dependence exists between Z_A and Z_B.

Depending on which of models A and B in (23) is used as the maintained hypothesis, the NT's, JT's and ET's give rise to cases where one either rejects or accepts both models. Table 1 gives four possible cases.

Before comparing the performances of a Bayesian MSEF criterion, the AIC, Efron's confidence interval, NT's, JT's, and ET's by sampling experiments, we need to develop a measure of distance between two competing

Table 1: Four Cases of NT's, JT's and ET's

	N_0, J_0, and E_0 Tests	
N_1, J_1, E_1 Tests	Case 1: Accept Model A Reject Model B (p_1)	Case 2: Reject Model A Accept Model B (p_2)
	Case 3: Reject Model A Reject Model B (p_3)	Case 4: Accept Model A Accept Model B (p_4)

Notes: p_i is the probability of case i, $(i = 1, \cdots, 4)$
NT's= N-tests, consisting of N_0 and N_1 tests
JT's=J-tests, consisting of J_0 and J_1 tests
ET's= encompassing tests, consisting of E_0 and E_1 tests

models. The distance measure should incorporate factors that influence the distance such as the correlations among non-overlapping explanatory variables of the two models (X_A and X_B in (23)), the parameters of the models, β_A, β_B and σ^2, and the dimensions of the models, d_A and d_B. It is also desirable that the distance measure is estimable from data. The difference of the MSE's of the two models, Δ, which Efron (1984) uses has an unbiased estimator [equation (24)] and it is a function of correlations among the regressors, the parameters of the models, and the dimensions of the two models.

Suppose that the dependent variable, y, in (23) is distributed as $y \sim N(\eta, \sigma^2 I_n)$. Then the MSE of model i MSE_i, (i =A, B) is

$$\begin{aligned} MSE_i &= E \mid y_i - \eta \mid^2 \\ &= \mid \eta - \eta_i \mid^2 + \sigma^2 d_i \quad i = A, B \end{aligned}$$

where y_i and η_i are the projections of y and η, respectively, into the space of model i. The difference of MSE_i's, $\Delta = MSE_B - MSE_A$, is given by

$$\begin{aligned} \Delta = {} & \eta' M_B^* X_A (X_A' M_B^* X_A)^{-1} X_A' M_B^* \eta \\ & - \eta' M_A^* X_B (X_B' M_A^* X_B)^{-1} X_B M_A^* \eta + (d_B - d_A)\sigma^2 \end{aligned} \tag{29}$$

If model A is true, *i.e.* $\eta = Z_A\delta_A$, then we have $M_A^*\eta = 0$, and Δ becomes

$$\Delta = \beta_A' X_A' M_B^* X_A \beta_A + (\mathrm{d}_B - \mathrm{d}_A)\sigma^2 \tag{30}$$

The first term on the right hand side may be reduced to

$$\beta_A' X_A' M_A^* X_A \beta_A = \begin{cases} (1-\rho_1^2)\lambda_1\zeta_1^2 + \cdots + (1-\rho_{\mathrm{d}_A}^2)\lambda_{\mathrm{d}_A}\zeta_{\mathrm{d}_A}^2, & \text{if } \mathrm{d}_A \le \mathrm{d}_B \\ (1-\rho_1^2)\lambda_1\zeta_1^2 + \cdots + (1-\rho_{\mathrm{d}_B}^2)\lambda_{\mathrm{d}_B}\zeta_{\mathrm{d}_B}^2 & \\ \quad + \lambda_{\mathrm{d}_B+1}\zeta_{\mathrm{d}_B+1}^2 + \cdots + \lambda_{\mathrm{d}_A}\zeta_{\mathrm{d}_A}^2 & \text{if } \mathrm{d}_A > \mathrm{d}_B \end{cases}$$

where $\zeta = (\zeta_1, \cdots, \zeta_{d_A})' = \mathrm{H}\beta_A$; H is a $\mathrm{d}_A \times \mathrm{d}_A$ orthogonal matrix; λ_i is the i-th characteristic root of $X_A' M_w X_A$ $(i = 1, \cdots, \mathrm{d}_A)$; ρ_i^2 is the i-th characteristic root of $(X_A' M_w X_A)^{-1} X_A' M_w X_B (X_B' M_w X_B)^{-1} X_B' M_w X_A$. ρ_i^2 can be interpreted as the i-th partial cannonical correlation between X_A and X_B after removing the influences of common regressors W. ζ is a function of β_A, the regression coefficients associated with the distinct regressors in the true model. Accordingly, the difference of MSE's, Δ, may be used as a measure of the distance between two models. If $X_A = X_B D$, where D is a $\mathrm{d}_B \times \mathrm{d}_A$ matrix of constants with rank $k = \min(\mathrm{d}_A, \mathrm{d}_B)$, then models A and B in (23) are the same. In this case, $\rho_1 = \cdots = \rho_k = 1$ and $\lambda_{d_B+1} = \cdots = \lambda_{d_A} = 0$ if $\mathrm{d}_B < \mathrm{d}_A$, and the difference of the MSE's, Δ, is given by $\Delta = (\mathrm{d}_B - \mathrm{d}_A)\sigma^2$. If one wishes to define the distance of two models to be zero when $X_A = X_B D$, then one may translate Δ by $(\mathrm{d}_B - \mathrm{d}_A)\sigma^2$ and define $\Delta^* = \Delta - (\mathrm{d}_B - \mathrm{d}_A)\sigma^2$. As $\rho_i^2 \to 0$ for $i = 1, \cdots, k$, Δ increases, and at $\rho_i^2 = 0$ $(i = 1, \cdots, k)$, Δ becomes $\Delta = \sum_{i=1}^{d_A} \lambda_i \zeta_i^2 + (\mathrm{d}_B - \mathrm{d}_A)\sigma^2$.

Sampling experiments are made first by specifying the two models as

$$\begin{aligned} \text{Model A: } y_t &= \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + u_t \\ \text{Model B: } y_t &= \gamma_0 + \gamma_1 x_{t1} + \gamma_2 z_{t2} + u_t \end{aligned} \tag{31}$$

Hence, models A and B have $(1, x_{t1})$ as the common variables, whereas x_{t2} and z_{t2} are uncommon variables. As in Pesaran's (1982) experiments, x_{ti}'s are drawn from $N(0,1)$, and z_{t2} is generated by

$$z_{t2} = \phi x_{t2} + v_{t2} \qquad v_{t2} \sim N(0,1) \tag{32}$$

and ϕ is chosen to make the simple population correlation of z_{t2} and x_{t2} equal to ρ. Hence,

$$\phi = \rho/(1-\rho^2)^{\frac{1}{2}}$$

where $\rho = \text{Corr}(x_{t2}, z_{t2})$.

The model selection criteria are also influenced by the fit of the true model as measured by the coefficient of determination of the true model (model A in our experiments), R^2 and by the size of β_2, the coefficient associated with the distinct regressor of the true model. We report the results of sampling experiments by setting R^2 at .5, .7, and .9 and β_1 and β_2 as 1.0 and .5, respectively, for most of the cases in Tables 2, 3, and 4. The constant term, β_0, is set at 1.0 in all the experiments. The number of replications for each value of ρ^2, R^2, and sample size is 500.

The following observations can be made from Tables 2,3, and 4.

(1) As the sample size increases the probabilities of choosing the correct model increase for all the criteria for given values of ρ^2, and R^2.

(2) As the coefficient of determination, R^2, increases the probabilities of choosing the correct model tends to increase for all the criteria for given values of ρ^2 and R^2.

(3) When ρ^2 is large the NT's tends to perform better than the JT's and ET's. For $\rho^2 = .1$ or 0, NT's probabilities of choosing the correct model decline. This is due to the fact that for low values of ρ^2, the non-overlapping regressors, x_{t2} and z_{t2} tend to be orthogonal, and this bring the NT's closer to the case in which it is not defined (*i.e.* $Z_A' Z_B = 0$).

Table 2: Empirical Probabilities of Choosing the Correct Model [$R^2 = .5$]

β_1	β_2	ρ^2	Δ	NT's $p_1^{(4)}$	JT's ET's $p_1^{(4)}$	Dif. of Post. Means[5]	Dif. of AIC[6]	Efron's 90% CI[7]
						n=20		
1.0	.5	1.0	0	ND[8]	ND[8]	1	1[9]	0[9]
1.0	.5	.9	.8	.216	.054	.84	.662	0
1.0	.5	.7	2.5	.402	.150	.932	.782	.002
1.0	.5	.5	3.8	.466	.238	.964	.828	.024
1.0	.5	.3	4.6	.668	.318	.978	.859	.028
1.0	.5	.1	5.1	.680	.308	.988	.872	.078
1.0	.5	0	5.9	.722	.418	.982	.848	.154
.5	1.0	0	23.6	.846	.906	1.0	1.0	.696
						n=60		
1.0	.5	1.0	0	ND	ND	1	1	0
1.0	.5	.9	1.7	.262	.182	.710	.752	0
1.0	.5	.7	5.2	.600	.392	.876	.884	.004
1.0	.5	.5	8.7	.770	.638	.936	.936	.114
1.0	.5	.3	12.2	.880	.772	.972	.982	.182
1.0	.5	.1	15.7	.548	.902	.990	.990	.574
1.0	.5	0	17.4	.812	.888	.994	.990	.642
.5	1.0	0	69.6	.876	.948	1.0	1.0	1.0
						n=100		
1.0	.5	1.0	0	ND	ND	1	1	0
1.0	.5	.9	2.5	.406	.288	.790	.800	0
1.0	.5	.7	7.8	.700	.570	.920	.916	0
1.0	.5	.5	13.7	.906	.834	.972	.962	.450
1.0	.5	.3	20.3	.940	.928	.992	.994	.702
1.0	.5	.1	27.9	.906	.946	1.0	1.0	.812
1.0	.5	0	32.7	.566	.952	1.0	1.0	.962
.5	1.0	0	130.8	.714	.954	1.0	1.0	1.0

Notes: (1) NT's=N-tests; JT's=J-tests; ET's=encompassing tests. ET's are identical to JT's for d_A=d_B=1

(2) For each value of ρ and n, the number of replications is 500.

(3) ρ = Corr(x_{t2}, z_{t2}); Δ is the measure of the distance between two models A and B [Δ in equation (29).] Given β_1 and β_2, Δ is a monotone function of ρ.

Footnotes of Table 2 continued:

(4) p_1 is the probability of accepting model A and rejecting model B. [See Table 1.]
(5) The posterior mean of $MSEF_i$ is computed by $E(MSEF_i|$ data) for each model, i =A, B using numerical integration by Simpson's rule, and the difference is $E(MSEF_B)$- $E(MSEF_A)$
(6) The difference of the AIC's is AIC_B-AIC_A.
(7) Efron's 90% confidence interval (CI) is computed by assuming that the sample estimate, $\tilde{\sigma}^2$, is true (hence, $d_E = \infty$ in Efron's notation), and by the Edgeworth expansion for the parametric bootstrap distribution without resorting to Monte Carlo. Monte Carlo generation of parametric bootstrap distributions yields the results quite close to those by the Edgeworth expansion.
(8) For $\rho^2 = 1$, the NT's, JT's, and ET's are not defined.
(9) For $\rho^2 = 1$, the difference of the posterior means of the MSE's and the difference of the AIC's both become 1 by construction. Efron's CI becomes zero by construction.
(10) For all the sample sizes, the periods of prediction, m is set at 10.

Table 3: Empirical Probabilities of Choosing the Correct Model [$R^2 = .7$]

β_1	β_2	ρ^2	Δ	NT's p_1	JT's ET's p_1	Dif. of Post. Means	Dif. of AIC	Efron's 90% CI
						n=20		
1.0	.5	1.0	0	ND	ND	1	1	0
1.0	.5	.9	.8	.386	.210	.620	.714	0
1.0	.5	.7	2.5	.717	.542	.810	.868	.018
1.0	.5	.5	3.8	.888	.684	.835	.952	.112
1.0	.5	.3	4.6	.828	.722	.835	.952	.247
1.0	.5	.1	5.1	.849	.768	.925	.960	.272
1.0	.5	0	5.9	.818	.738	.953	.956	.438
.5	1.0	0	23.6	.868	.945	1.0	1.0	.990
						n=60		
1.0	.5	1.0	0	ND	ND	1	1	0
1.0	.5	.9	1.7	.528	.426	.802	.774	.002
1.0	.5	.7	5.2	.908	.866	.920	.950	.424
1.0	.5	.5	8.7	.928	.946	.973	.990	.746
1.0	.5	.3	12.2	.914	.942	.994	.998	.914
1.0	.5	.1	15.7	.874	.948	.999	1.0	.967
1.0	.5	0	17.4	.868	.948	1.0	1.0	.970
.5	1.0	0	69.6	.969	.945	1.0	1.0	1.0
						n=100		
1.0	.5	1.0	0	ND	ND	1	1	0
1.0	.5	.9	2.5	.652	.559	.871	.958	.007
1.0	.5	.7	7.8	.923	.877	.950	.941	.584
1.0	.5	.5	13.7	.951	.955	.987	.977	.844
1.0	.5	.3	20.3	.932	.951	.999	1.0	.988
1.0	.5	.1	27.9	.915	.960	1.0	1.0	.996
1.0	.5	0	32.7	.862	.945	1.0	1.0	1.0
.5	1.0	0	130.8	.912	.945	1.0	1.0	1.0

See footnotes below Table 2.

Table 4: Empirical Probabilities of Choosing the Correct Model [$R^2 = .9$]

β_1	β_2	ρ^2	Δ	NT's p_1	JT's ET's p_1	Dif. of Post. Means	Dif. of AIC	Efron's 90% CI
						n=20		
1.0	.5	1.0	0	ND	ND	1	1	0
1.0	.5	.9	.8	.784	.588	.771	.884	.012
1.0	.5	.7	2.5	.910	.938	.913	.988	.401
1.0	.5	.5	3.8	.930	.970	.965	1.0	.556
1.0	.5	.3	4.6	.886	.980	.998	1.0	.875
1.0	.5	.1	5.1	.638	.950	.998	1.0	.899
1.0	.5	0	5.9	.886	.945	1.0	1.0	.988
.5	1.0	0	23.6	.936	.945	1.0	1.0	1.0
						n=60		
1.0	.5	1.0	0	ND	ND	1	1	0
1.0	.5	.9	1.7	.914	.873	.957	.990	.168
1.0	.5	.7	5.2	.947	.949	.996	.993	.956
1.0	.5	.5	8.7	.939	.936	1.0	1.0	1.0
1.0	.5	.3	12.2	.935	.958	1.0	1.0	1.0
1.0	.5	.1	15.7	.939	.958	1.0	1.0	1.0
1.0	.5	0	17.4	.928	.946	1.0	1.0	1.0
.5	1.0	0	69.6	.936	.945	1.0	1.0	1.0
						n=100		
1.0	.5	1.0	0	ND	ND	1	1	0
1.0	.5	.9	2.5	.943	.945	.993	1.0	.655
1.0	.5	.7	7.8	.944	.949	1.0	1.0	.916
1.0	.5	.5	13.7	.959	.963	1.0	1.0	1.0
1.0	.5	.3	20.3	.944	.951	1.0	1.0	1.0
1.0	.5	.1	27.9	.938	.960	1.0	1.0	1.0
1.0	.5	0	32.7	.911	.945	1.0	1.0	1.0
.5	1.0	0	130.8	.912	.945	1.0	1.0	1.0

See footnotes below Table 2.

(4) The performances of the Bayesian MSEF and AIC criteria are similar. Both of them have higher probabilities of choosing the correct model than the other criteria. In the case of R^2=.5 with sample size of 20, the Bayesian criterion has higher probabilities of choosing the correct model than the AIC.

(5) Efron's 90% confidence interval (CI) appears to be quite conservative when sample sizes are small and when ρ^2 is close to zero. For higher R^2's and higher ρ^2, however, the performances of Efron's 90% CI's improve markedly. As shown in Efron (1984) the probability that the 90% CI contains $\Delta = 0$ is larger than 90%, indicating that it is harder to reject $\Delta = 0$.

(6) For $\rho^2 = 0$, we present two results: one by setting (β_1, β_2)=(1, .5) and the other by setting (β_1, β_2)=(.5, 1). These results are presented to show that the distance between the two models, Δ, is a function of β_2. As one expects from the definition of Δ in equation (29), $\beta_2 = 1$ yields larger values of Δ than $\beta_2 = .5$ does even if $\rho^2 = 0$ holds for both cases. The higher values of Δ produce higher probabilities of choosing the correct model.

Efron (1984) suggests a non-parametric bootstrapping procedure in addition to a parametric bootstrapping procedure. Since the non-parametric bootstrapping procedure requires considerable computational time in conducting sampling experiments, we did not carry it out in our experiments. However, for the case of R^2=.5, $n = 20$ and $n = 60$, we tried non-parametric bootstrapping procedures, and we found that confidence intervals that are generated by non-parametric bootstrapping tend to be larger than those by parametric bootstrapping, and thus the probabilities of choosing the correct model are in general lower than those by parametric bootstrapping.

In our sampling experiments, we set the prediction period, m, at 10. We varied m at different values, and the results are comparable to those of m=10.

The models in equation (31) assume that d_A=d_B. Will the results in Tables 2, 3, and 4 be sensitive to whether we have $d_A > d_B$ or $d_A < d_B$? Suppose that the true model, model A, is the same as in (31) but model B

has $d_B > d_A$:

$$\begin{aligned} \text{Model A: } y_t &= \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + u_t \\ \text{Model B: } y_t &= \gamma_0 + \gamma_1 x_{t1} + \gamma_2 z_{t2} + \gamma_3 z_{t3} + \gamma_4 z_{t4} + u_t \end{aligned} \tag{33}$$

In equation (33) we have $d_A = 1 < d_B = 3$. As before, x_{ti} $(i = 1, 2)$ are drawn from $N(0,1)$ and z_{tj} $(j = 2, 3, 4)$ are generated by

$$z_{tj} = \phi\, x_{t2} + v_{tj} \quad v_{tj} \sim N(0,1)$$

and ϕ is determined by the correlation between z_{tj} and x_{t2}:

$$\phi = \rho/(1-\rho^2)^{1/2}$$

where ρ =Corr(x_{t2}, z_{tj}), $j = 2, 3, 4$.

The results of sampling experiments for $R^2 = .7$ are given in Table 5 for $n = 20$, 60, and 100. Comparing Tables 3 and 5, we see that in general the results in these two tables are similar. We reported here only for $R^2 = .7$ since the results for $R^2 = .5$ and .9 are comparable to those for $R^2 = .7$ except that as R^2 increases the probabilities of choosing the correct model increase for all the model selection criteria.

Let us change the role of model A and B in (33):

$$\begin{aligned} \text{Model A: } y_t &= \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + u_t \\ \text{Model B: } y_t &= \gamma_0 + \gamma_1 x_{t1} + \gamma_2 z_{t2} + u_t \end{aligned} \tag{34}$$

We now have $d_A = 3 > d_B = 1$. As before, x_{ti} $(i = 1, \cdots, 4)$ are drawn from $N(0,1)$ and z_{t2} is generated by

$$z_{t2} = \phi x_{t2} + v_{t2}, \quad v_t \sim N(0,1)$$

The results of sampling experiments for $R^2 = .7$ are given in Table 6 for $n = 20, 60$, and 100 and $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = (1,\ 1,\ .5,\ .05,\ .05)$. Comparing Table 5 and Table 6, we find that for sample size of 20, the performance of the Bayesian MSEF criterion is noticably poorer when $d_A > d_B$. This is due to the fact that we use 10 observations for estimation and 10 observations for post-sample forecast thus leaving not enough observations for estimating

Table 5: Empirical Probabilities of Choosing the Correct Model When $d_A = 1 < d_B = 3$ [$R^2 = .7$]

ρ^2	Δ^*	NT's p_1	JT's p_1	ET's p_1	Dif. of Post. Means	Dif. of AIC	Efron's 90% CI
				n=20			
1.0	0	ND	ND	ND	1	1	0
.9	.14	.106	.06	.048	.972	.920	0
.7	.46	.258	.130	.116	.984	.926	.024
.5	.89	.406	.224	.196	1.0	.928	.070
.3	1.53	.578	.367	.346	1.0	.924	.112
.1	2.69	.676	.500	.502	1.0	.954	.226
0	4.21	.678	.614	.708	1.0	.984	.372
				n=60			
1.0	0	ND	ND	ND	1	1	0
.9	.57	.204	.174	.164	.992	.944	.012
.7	2.02	.554	.424	.428	.996	.944	.172
.5	4.06	.818	.754	.754	1.0	.974	.412
.3	7.16	.892	.890	.912	1.0	.996	.624
.1	12.41	.862	.910	.960	1.0	1.0	.890
0	16.22	.464	.738	.946	1.0	1.0	.982
				n=100			
1.0	0	ND	ND	ND	1	1	0
.9	.9	.280	.236	.236	.968	.942	.024
.7	2.9	.684	.608	.604	.972	.926	.270
.5	5.8	.864	.846	.860	.988	.972	.496
.3	10.5	.900	.906	.926	.972	.996	.742
.1	19.5	.908	.942	.956	1.0	1.0	.950
0	36.6	.800	.850	.956	1.0	1.0	.998

Notes: (1) See the footnotes in Table 2. All the experiments are made by setting $\beta_0 = \beta_1 = 1$ and $\beta_2 = .5$.

(2) Δ^* is the distance between the two models that is adjusted to make the distance zero when $\rho^2 = 1.0$ *i.e.* $\Delta^* = \Delta - (d_B - d_A)\sigma^2$.

Table 6: Empirical Probabilities of Choosing the Correct Model When $d_A = 3 > d_B = 1$ [$R^2 = .7$]

ρ^2	Δ^*	NT's p_1	JT's p_1	ET's p_1	Dif. of Post. Means	Dif. of AIC	Efron's 90% CI
					n=20		
1.0	0	ND	ND	ND	1	1	0
.9	.6	.626	.394	.098	.112	.232	.072
.7	1.6	.754	.542	.150	.368	.342	.088
.5	2.8	.844	.710	.350	.556	.600	.204
.3	4.1	.894	.836	.500	.704	.746	.256
.1	5.7	.838	.902	.620	.796	.826	.392
0	6.8	.616	.902	.648	.822	.882	.400
					n=60		
1.0	0	ND	ND	ND	1	1	0
.9	2.2	.748	.692	.348	.896	.528	.208
.7	6.1	.926	.932	.770	.996	.868	.500
.5	10.1	.940	.934	.898	1.0	.966	.724
.3	14.0	.936	.946	.936	1.0	.984	.904
.1	17.1	.872	.944	.942	1.0	.996	.928
0	17.2	.840	.924	.938	1.0	1.0	.932
					n=100		
1.0	0	ND	ND	ND	1	1	0
.9	2.6	.734	.690	.382	.752	.564	.244
.7	6.7	.940	.934	.814	.972	.880	.544
.5	11.2	.940	.966	.948	.996	.974	.708
.3	16.6	.950	.962	.960	1.0	.994	.928
.1	23.9	.946	.966	.966	1.0	1.0	.984
0	31.1	.910	.960	.960	1.0	1.0	1.0

Notes: (1) See the footnotes in Table 5.
(2) Model A in equation (34) is the true model.
(3) Parameters in Model A: $(\beta_0, \beta_1, \beta_3, \beta_4) = (1, 1, .5, .05, .05)$

the models. Also, we see that the performance of the AIC is noticeably poorer when $d_A > d_B$ than when $d_A < d_B$, regardless of sample sizes of 20, 60, and 100. This is because the penalty factor for employing a larger number of regressors ($2k$ in equation (22)) tends to penalize excessively the true model that has a larger number of regressors. The penalty term of the AIC has been discussed in Sawa (1978), Schwarz (1978), and Chow (1983) who make adjustments to the AIC.

How will the model selection criteria perform if neither of models A and B is true ? Suppose that models A and B are given as in (31) and they are considered for selection, but y_t is generated by $N(\eta_t, \sigma^2)$, where η_t is given by $\eta_t = \alpha_0 + \alpha_1 x_{t1} + \alpha_2 w_{t2}$. Hence the true model is

$$\text{Model C:} \quad y_t = \alpha_0 + \alpha_1 x_{t1} + \alpha_2 w_{t2} + u_t \tag{35}$$

In data analysis, a researcher seldom knows the true model that generates the data to be analyzed. Many of those who have proposed different model selection criteria are aware of a possibility that models may not be the true model. The proponents of the model selection criteria that are based on goodness of fit intend to seek a model which explains data better than any other model. The proponents of the model selection criteria that are made in hypotheis testing framework are also aware of the fact that none of the models may be correct. In proposing the encompassing principle, Mizon and Richard (1986) state a need to search for useful models that are only an approximation of the true model. What one can hope for, then, is that a model selection criterion tends to choose a model that is closer to the true model.

We let model A in (31) be always closer to the true model C in (35) by making the correlation between x_{t2} and w_{t2} higher than that between z_{t2} and w_{t2}. Specifically, we draw x_{ti}, $(i = 1, 2)$, w_{t2}, and z_{t2} as follows:

$$\begin{aligned} x_{t1},\ w_{t2} &\sim N(0,1) \\ x_{t2} &= \left[\rho_{xw}/(1-\rho_{xw}^2)^{1/2}\right] w_{t2} + v_{t2} \\ z_{t2} &= \left[\rho/(1-\rho^2)^{1/2}\right] x_{t2} + v_{t3} \\ v_{tj} &\sim N(0,1),\ \ j = 2,3 \end{aligned}$$

where ρ_{xw} =Corr(x_{t2}, w_{t2}), and ρ = Corr(x_{t2}, z_{t2}). We hold ρ^2_{xw} at .9 and vary ρ^2 between 0 and 1, and set $(\alpha_0, \alpha_1, \alpha_2) = (1, 1, .5)$. Table 7 gives the results for $R^2 = .7$ and $n = 60$. Under each test criterion there are two subcategories (I) and (II). Category (I) is the case where model A in (31) is correct, and hence p_1 columns for NT's and JT's (ET's) are the same as those in Table 3 for $n = 60$. The numbers under (I) in Dif of Post. Means and in Dif of AIC's are the same as those in Table 3. Category (II) is the case where model C in (35) is correct. From Table 7, we may observe

(1) The performances of the Bayesian MSEF and AIC criteria are slightly reduced when model C is correct, but these criteria tend to choose model A that is closer to the true model (model C).

(2) As the correlations between x_{t2} and z_{t2} in models A and B increase, the probabilities of accepting both models A and B increase for the NT's, JT's and ET's when both of models A and B are wrong.

(3) As the correlation between x_{t2} and z_{t2} in models A and B increase, the probabilities of accepting the null hypothesis that models A and B are not different increase for Efron's 90% confidence interval.

So far we have presented the results of sampling experiments. Here let us apply the Bayesian MSEF and other criteria to Williams' data (1959) which Efron (1984) uses as an example to illustrate his confidence interval criterion. Williams two models are

$$\begin{aligned} \text{Model A: } y_i &= \beta_0 + \beta_1 x_i + u_i \\ & \qquad\qquad\qquad\qquad i = 1, \cdots, n \\ \text{Model B: } y_i &= \gamma_0 + \gamma_1 z_i + u_i \end{aligned}$$

where y_i = the maximum compression strength of a piece of wood, $lb/sq.in$ x_i = density of a piece of wood, $lb/cu.ft$, and z_i = adjusted density for a piece of wood, $lb/cu.ft$.

Table 7: Comparison of Propabilties of Model Selection: (I) Model A in (31) is Correct; (II) Model C in (35) is Correct and Model A is closer to Model C than Model B [$R^2 = .7$, $n = 60$]

			NT's					
			(I)			(II)		
ρ^2	Δ_I	Δ_{II}	p_1	p_3	p_4	p_1	p_3	p_4
1.0	0	0	ND	ND	ND	ND	ND	ND
.9	1.7	.26	.528	.018	.420	.098	.034	.848
.7	5.2	.35	.908	.046	.022	.272	.026	.670
.5	8.7	.84	.928	.070	.000	.476	.026	.670
.3	12.2	1.79	.914	.086	.000	.784	.020	.170
.1	15.7	4.75	.874	.126	.000	.962	.060	.000
0	17.4	11.74	.868	.132	.000	.760	.240	.000
			JT's and ET's					
			(I)			(II)		
ρ^2	Δ_I	Δ_{II}	p_1	p_3	p_4	p_1	p_3	p_4
1.0	0	0	ND	ND	ND	ND	ND	ND
.9	1.7	.26	.426	.022	.527	.068	.04	.876
.7	5.2	.35	.866	.030	.084	.202	.030	.750
.5	8.7	.84	.946	.040	.010	.384	.016	.588
.3	12.2	1.79	.942	.058	.000	.650	.032	.306
.1	15.7	4.75	.948	.052	.000	.930	.040	.028
0	17.4	11.74	.949	.058	.000	.962	.038	.000
			Efron's 90% CI					
			(I)			(II)		
ρ^2	Δ_I	Δ_{II}	$\Delta > 0$	$\Delta < 0$	$\Delta \simeq 0$	$\Delta > 0$	$\Delta < 0$	$\Delta \simeq 0$
1.0	0	0	0	0	1	0	0	1
.9	1.7	.26	.002	0	.998	0	0	1
.7	5.2	.35	.424	0	.576	0	0	1
.5	8.7	.84	.746	0	.254	0	0	1
.3	12.2	1.79	.914	0	.086	.056	0	.944
.1	15.7	4.75	.964	0	.036	.672	0	.328
0	17.4	11.74	.970	0	.030	.914	0	.086
			Dif of Post Means			Dif of AIC's		
ρ^2	Δ_I	Δ_{II}	(I)		(II)	(I)		(II)
1.0	0	0	1		1	1		1
.9	1.7	.26	.802		.854	.774		.674
.7	5.2	.35	.920		.894	.950		.770
.5	8.7	.84	.973		.922	.990		.808
.3	12.2	1.79	.994		.950	.998		.912
.1	15.7	4.75	.999		.984	1		.998
0	17.4	11.74	1		.994	1		1

Notes to Table 7

(1) $\Delta_I = \text{MSE}_B - \text{MSE}_A$ under (I), $\Delta_{II} = \text{MSE}_B - \text{MSE}_A$ under (II).

(2) p_1= Probability of accepting Model A and rejecting Model B.
p_3= Probability of rejecting both Model A and Model B.
p_4= Probability of accepting both Model A and Model B.

(3) $\Delta = \text{MSE}_B - \text{MSE}_A$. The numbers under $\Delta > 0$ show the empirical probabilities that Efron's 90% CI criterion chooses Model A over Model B. The numbers under $\Delta < 0$ show the empirical probabilities that Efron's 90% CI criterion chooses Model B over model A, and the numbers under $\Delta = 0$ show the empirical probabilities that Efron's 90% CI criterion chooses neither Model A nor Model B.

We used 21 of 42 observations for estimating the models and 21 obseervations for post-sample forecast (hence m=21). The summary statistics of the posterior pdf's of the two MSEF's are given in Table 8 along with the statistics of the other model selection criteria. The difference of the posterior means of the MSEF's shows that model B should be preferred to model A. The 95% highest posterior density intervals (HPDI's) show that the interval is much tighter for model B than for model A. The difference of the AIC's also leads to the choice of model B, and so does Efron's 90% confidence interval. The NT's, JT's and ET's also lead to the acceptance of model B and rejection of model A. All the model selection criteria, thus, lead to the same conclusion to choose model B over model A. This is somewhat expected from the results of our sampling experiments. Tables 2, 3, 4, and 7 show that when Δ/σ^2 (in our experiments, $\sigma^2 = 1$) is over 17, the probabilities of choosing the correct model or the model that is closer to the unknown true model) increase considerably for all the model selection criteria. Since the scaled Δ is estimated to be -20.1 we expect that Williams's data will lead to the same conclusion regardless of which criterion is used.

As indicated by the 95% HPDI's for the MSEF's, model B produces the posterior pdf that is much tighter than that for model A. Inspection of the posterior pdf's gives us more information than what is available from such point estimates as the posterior means, modes, and variances, and this is an attractive feature of the Bayesian procedure.

Table 8: Summary Statistics of the Model Selection Criteria for Williams's Data

	Model A	Model B
Bayesian MSEF's		
Posterior Mean	92,084	72.414
Mode	65,753	54,794
95% HPDI	(0, 180821)	(0, 142465)
AIC	336,456	319.398
NT's	$N_0 = -5.8197$	$N_1 = -1.2013$
JT's (=ET's)	$J_0 = 4.6099$	$J_1 = 1.068$
Efron's 90% CI for δ	$(-38.90, -1.74)$	
$\hat{\Delta}$	$-153,310$	
$\hat{\Delta}/\bar{\sigma}^2$	-20.11	
r_{xz}	.96	

Notes: $\delta = \Delta/\sigma$. Efron's confidence interval is obtained using $\bar{\sigma}$
$\bar{\sigma}$= estimate of σ using $y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i + u_i$.
$\hat{\Delta}$ = unbiased estimate of $\Delta = \text{MSE}_B - \text{MSE}_A$.
r_{xz}= sample correlation coefficient between the non-overlapping regressors x_i and z_i.

4. Conclusions

In this paper we derived Bayesian and sampling distributions for the MSEF of a linear regression model. We used a degenerate hypergeometric function rather than Laguerre expansions since the former appears to be easily adopted to numerical computation than the latter. We suggested that posterior statistics, such as the posterior means or highest posterior density intervals (HPDI's) of the MSEF's may be used as a model selection criterion. Using sampling experiments we evaluated the Bayesian MSEF criterion along with the AIC, Efron's confidence interval, NT's, JT's and ET's. We chose the difference of MSE's of two models, Δ, as the measure of distance between two models. This distance measure reflects correlations among non-overlapping regressors in two models, regression coefficient, the variance of the error terms, and dimensions of two models, and it has an unbiased estimator.

Our sampling experiments show that the Bayesian MSEF criterion performs as well as or sometimes better than the AIC, except in the case of a small sample size with the dimension of the true model exceeding that of the competing model. In the case of a small sample size one may be adviced to use as many observations as one can for the estimation of the parameters of the models, leaving only one or two observations for post-sample forecast. The AIC tends to penalize the true model that contains more parameters than the competing model, raising an issue about the penalty factor $2k$ in the AIC, the issue that has been discussed in the literature.

Efron's CI tends to be more conservative than the other criteria when the distance between two competing models is not large. We have adopted a strategy to reject model B in favor of model A if the 90% confidence interval for $\Delta = \text{MSE}_B - \text{MSE}_A$ lies entirely to the right of 0. When Δ is small, this strategy does not work well due to the fact that actual coverage propabilities that the 90% confidence interval contains $\Delta = 0$ is considerably greater than .9 as shown in Table 9. Table 9 gives coverage probabilities of 90% confidence intervals for the models given in equation (31). Efron (1984) notes in Table 4 of his article that in the simple case the confidence interval method works poorly near the origin. If one knows that Δ is small, then, one may loosen the level of confidence interval to adjust for high coverage probabilities of the confidence interval near $\Delta = 0$.

Table 9: Probabilties that Efron's 90% CI Contains $\Delta = \text{MSE}_B - \text{MSE}_A$: Models A and B in (31)

		$n = 20$		
ρ^2	Δ	$R^2 = .5$	$R^2 = .7$	$R^2 = .9$
1.0	0	1.0	1.0	1.0
.9	.8	1.0	.998	.920
.7	2.5	.920	.896	.838
.5	3.8	.868	.866	.786
.3	4.6	.854	.882	.780
.1	5.1	.860	.862	.760
0	5.9	.876	.888	.764
		$n = 60$		
ρ^2	Δ	$R^2 = .5$	$R^2 = .7$	$R^2 = .9$
1.0	0	1.0	1.0	1.0
.9	1.7	1.0	.948	.938
.7	5.2	.944	..944	.970
.5	8.7	.906	.892	.892
.3	12.2	.890	.884	.814
.1	15.7	.910	.890	.780
0	17.4	.890	.872	.762
		$n = 100$		
ρ^2	Δ	$R^2 = .5$	$R^2 = .7$	$R^2 = .9$
1.0	0	1.0	1.0	1.0
.9	2.5	1.0	.950	.948
.7	7.8	.950	.944	.892
.5	13.7	.936	.932	.862
.3	20.3	.902	.872	.792
.1	27.9	.900	.860	.732
0	32.7	.890	.870	.762

Notes: (1) Model A is correct.
(2) In each case, 500 replications are made.

When neither of two competing models is correct, Cox's separate families tests (NT's, JT's and ET's) tend to accept both models when the distance between these two models is not too large, while Efron's CI tends to accept the null hypothesis that the two models are not different from each other.

References

(1) Akaike, H. (1974). "A new look at the statistical model identification," *IEEE Transactions in Automatic Control*, AC-19, 716-723

(2) Bhattacharyya, A. (1945). "A note on the distribution of the sum of chi- squares," *Sankhya*, 7, 27-28

(3) Bhattacharyya, B.C. (1943). "On an aspect of the Pearson system of curves and a few analogies," *Sankya*, 6, 415-448

(4) Chow, G.C. (1983). *Econometrics*, McGraw-Hill, New York

(5) Davidson, R. and J.G. MacKinnon (1981). "Several tests for model specification in the presence of alternative hypotheses," *Econometrica* , 49, 781-793

(6) Efron, B. (1984). "Comparing non-nested linear models," *Journal of the American Statistical Association*, 79, 791-803

(7) Gradshteyn, I.S. and I.M. Ryzhik (1980). *Tables of integrals, series, and products* Academic Press, New York

(8) Gurland, J. (1953). "Distribution of quadratic forms and ratios of quadratic forms," *Annals of Mathematical Statistics,* 24, 416-427

(9) Hotelling, H. (1940). "The selection of variates for use in prediction with some comments on the general problem of nuisance parameters," *Annals of Mathematical Statistics*, 11, 271-283

(10) Hotelling, H. (1948). "Some new methods for distributions of quadratic forms," (Abstract), *Annals of Mathematical Statistics*, 19, 119

(11) Johnson, N.L. and S. Kotz (1970). *Continuous univariate distributions-2* , John Wiley and Sons, New York

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

(13) Mizon, G.E. and J.F. Richard (1986). "The encompassing principle and its application to testing non-nested hypotheses," *Econometrica*, 54, 657-678

(14) Neumann, J. von. (1941). "Distribution of the ratio of the mean square successive difference to variance," *Annals of Mathematical Statistics,* 12, 367-395

(15) Pesaran, M.H. (1974) "On the general problem of model selection," *Review of Economic Studies*, 41, 153-171

(16) Pesaran, M.H. (1982) "Comparison of local power of alternative tests of non-tested regression models," *Econometrica*, 50, 1287-1304

(17) Sawa, T. (1978). "Information criteria for discriminating among alternative regression models," *Econometrica*, 46, 1273-1291

(18) Schwarz, G. (1978). "Estimating the dimension of a model," *Annals of Statistics*, 6, 461-464

(19) Williams, E. (1959). *Regression analysis*, Johsn Wiley and Sons, New York

(20) Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*, John Wiley and Sons, New York