The Distance Between Regression Models and Its Impact on Model Selection

Philip Hans Franses*

Econometric Institute and Tinbergen Institute
Erasmus University Rotterdam
P.O. Box 1738
NL-3000 DR Rotterdam, The Netherlands

Transmitted by John Casti

ABSTRACT

In the case of nested hypothesis testing there exist many model selection criteria with which a "best" model is chosen. It is however very likely that the estimated hypotheses are all "incorrect." In this paper model choice is made while assuming the existence of a "true" model. After the introduction of the distance as a measure of difference between models, it is demonstrated that model selection is influenced by the correct specification. Eight choice criteria are compared with respect to this influence.

1. INTRODUCTION

In nested hypothesis testing there exist many model selection criteria, and in the econometric and statistical literature several articles deal with a comparison of these rules. The determination of the penalties the criteria impose on the inclusion of regressors (see e.g. [10]) is an important aspect in this comparison. In Engle and Brown [4] several selection criteria are reviewed, and their behavior in a forecasting experiment is studied. It is one of their conclusions that the criteria with the larger penalties on the number

*A previous draft of this paper was presented at the European Meeting of the Econometric Society 1988 in Bologna, Italy. Many helpful comments from Teun Kloek are gratefully acknowledged. All errors are my own.

655 Avenue of the Americas, New York, NY 10010 0096-3003/89/$03.50

of parameters perform better in selecting forecasting models. In [4] the selection rules are functions of the estimated residual sums of squares. Based on the choice criteria, a certain "best" model is chosen.[1] Nevertheless, it is very likely that the estimated models are all "wrong." In this paper the choice is made between two estimated incorrect models while assuming the existence of a "true" model. It is demonstrated that this correct specification has an effect on the model selection.

In Section 2 we introduce the concept of distance as a measure of difference between regression models. In Sneek [10] an analogous concept has already been introduced for the difference between ARMA models, and it has been seen as useful for evaluating alternative order selection rules in simulation experiments. In our paper, it is analytically shown that a choice rule for regression models can be expressed in terms of, first, the distance between the two estimated models and, second, a critical value. Both terms depend on the correct model, and hence the hypothesized true specification can influence the choice. In Section 3 eight selection criteria are discussed and compared. In Section 4 a small experiment is carried out to illustrate the obtained analytical results. The last section consists of the discussion and concluding remarks.

2. TWO CONCEPTS IN THE NESTED HYPOTHESIS TESTING FRAMEWORK

Consider the nested hypotheses

$$H_1: y = X_1\beta_1 + X_2\beta_2 + \epsilon, \tag{2.1}$$

$$H_2: y = X_1\beta_1 + \epsilon, \tag{2.2}$$

where y is a $T \times 1$ vector, X_1 a $T \times k_1$ matrix, X_2 a $T \times k_2$ matrix, β_1 a $k_1 \times 1$ vector, β_2 a $k_2 \times 1$ vector, and ϵ a $T \times 1$ vector. It is assumed that the disturbances are independently and identically distributed with zero mean and variance σ^2. The vector ϵ is defined as $y - E(y|X)$, where X is the matrix of observations for exogenous variables. Furthermore, we suppose the existence of a "true" model, in which the endogenous variable y is generated

[1] In this paper such expressions as "best" and "true" models are used without further explanation. I am aware of their questionable status, especially considering developments in the methodology of econometric model building during the past decade. For the analysis presented here they are viewed as only hypothetical, though useful, constructs.

by

$$H_t: y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \epsilon, \tag{2.3}$$

where X_3 is a $T \times k_3$ matrix, β_3 is a $k_3 \times 1$ vector, and $k_3 \neq 0$.

For notational convenience we write the disturbances in the three models as ϵ_1, ϵ_2, and ϵ respectively. Noting that $\epsilon_1 = X_3\beta_3 + \epsilon$ and $\epsilon_2 = X_2\beta_2 + X_3\beta_3 + \epsilon$, we compare the residual variances of the estimated "wrong" models H_1, H_2 and the "true" model H_t. The residual variance of H_1 is denoted by

$$s^2 = \frac{y'My}{T}, \tag{2.4}$$

where $M = I - X(X'X)^{-1}X'$ and X is $(X_1 \vdots X_2)$. The variable y is, however, generated by (2.3), and substituting this expression in (2.4) gives

$$s^2 = \frac{\beta_3' X_3' M X_3 \beta_3}{T} + \frac{2\beta_3' X_3' M \epsilon}{T} + \frac{\epsilon' M \epsilon}{T} \tag{2.5}$$

because of the properties of M: $M' = M$, $M^2 = M$, and $MX = 0$. When we denote t^2 as the residual variance belonging to (2.3), it can be shown (see Kloek [6]) that $t^2 < s^2$ with a probability close to 1 provided that the sample size is large enough.

The second term of the right hand side of (2.5) converges to zero in the squared mean (see [6]). Furthermore, it is assumed that $\text{plim}(X'X)/T$ is nonstochastic, finite, and of full rank, and that when $T \to \infty$, $\text{plim}\, X'\epsilon/\sqrt{T} \to 0$ (see [5, Chapter 5] for a complete survey of the assumptions). Together this implies that

$$\text{plim}\, s^2 = \text{plim} \frac{\beta_3' X_3' M X_3 \beta_3}{T} + \sigma^2 \tag{2.6}$$

and

$$\text{plim}\, t^2 = \sigma^2.$$

The difference between $\text{plim}\, s^2$ and $\text{plim}\, t^2$ is taken as a measure of distance between the models H_1 and H_t. For H_2 this can be established analogously.

2.1. The Distance

If we define $\hat{\epsilon}_1$ as $y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2$, or briefly $y - X\hat{\beta}$, then we define the distance D_1 between H_1 and H_t as T times the estimated distance[2]:

$$\begin{aligned} D_1 &= (\hat{\epsilon}_1 - \epsilon)'(\hat{\epsilon}_1 - \epsilon) \\ &= [y - X\hat{\beta} - (y - X\beta - X_3\beta_3)]'[y - X\hat{\beta} - (y - X\beta - X_3\beta_3)]. \end{aligned} \tag{2.7}$$

Under the true specification (2.3) we can write

$$X\hat{\beta} = X(X'X)^{-1}X'y = X\beta + PX_3\beta_3 + P\epsilon, \tag{2.8}$$

where $P = X(X'X)^{-1}X'$, and $P = I - M$. This symmetric and idempotent projection matrix P has the property that $MP = (I - P)P = 0$. When (2.8) is substituted in (2.7) this gives

$$\begin{aligned} D_1 &= (MX_3\beta_3 - P\epsilon)'(MX_3\beta_3 - P\epsilon) \\ &= \beta_3'X_3'MX_3\beta_3 + \epsilon'P\epsilon. \end{aligned} \tag{2.9}$$

The first derivative of D_1 with respect to $X_3\beta_3$ is

$$\frac{\partial D_1}{\partial X_3\beta_3} = 2MX_3\beta_3. \tag{2.10}$$

A property of idempotent and symmetric matrices is that they are positive semidefinite (p.s.d.). Moreover, if we assume that X_3 cannot be written as a linear combination of X_1 and/or X_2, the optimum value for D_1 will be reached when $X_3\beta_3 = 0$. The second derivative of D_1 is

$$\frac{\partial^2 D_1}{\partial X_3\beta_3\, \partial(X_3\beta_3)'} = 2M. \tag{2.11}$$

[2]The distances between the models H_1 and H_t, H_2 and H_t, and H_1 and H_2, denoted as D_1, D_2, and D respectively, are all defined as T times the respective estimated distances. Assuming a finite sample, this has no effect on the content of Sections 2 and 3, so we omit this factor T in further expressions for notational convenience.

This is a p.s.d. matrix, and thereby the established optimum in (2.10) is a minimum.

The distance between the estimated model H_2 and H_t is defined, analogously to (2.7), as

$$D_2 = (\hat{\epsilon}_2 - \epsilon)'(\hat{\epsilon}_2 - \epsilon), \tag{2.12}$$

where $\hat{\epsilon}_2 = y - X_1\hat{\beta}_1$. The right hand side of (2.12) can be written as

$$(-X_1\hat{\beta}_1 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3)'(-X_1\hat{\beta}_1 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3). \tag{2.13}$$

When

$$\begin{aligned} X_1\hat{\beta}_1 &= X_1(X_1'X_1)^{-1}X_1'y \\ &= X_1\beta_1 + P_1X_2\beta_2 + P_1X_3\beta_3 + P_1\epsilon, \end{aligned} \tag{2.14}$$

where $P_1 = X_1(X_1'X_1)^{-1}X_1' = I - M_1$ is substituted in (2.13), it yields

$$[M_1(X_2\beta_2 + X_3\beta_3) - P_1\epsilon]'[M_1(X_2\beta_2 + X_3\beta_3) - P_1\epsilon]. \tag{2.15}$$

Using that $M_1P_1 = M_1(I - M_1) = M_1 - M_1^2 = 0$, (2.12) becomes

$$D_2 = (X_2\beta_2 + X_3\beta_3)'M_1(X_2\beta_2 + X_3\beta_3) + \epsilon'P_1\epsilon. \tag{2.16}$$

The first derivative of D_2 with respect to $X_3\beta_3$ is

$$\frac{\partial D_2}{\partial X_3\beta_3} = 2M_1(X_2\beta_2 + X_3\beta_3). \tag{2.17}$$

The inverse matrix of M_1 does not exist. For determining the optimum value for D_2, we have to assume that the matrices X_2 and X_3 are of full rank, i.e., they are not linear combinations of the columns of X_1 and of each other. In this case, the optimum will be reached when $X_3\beta_3 = -X_2\beta_2$. This optimum is a minimum because the second derivative of D_2 is

$$\frac{\partial^2 D_2}{\partial X_3\beta_3\,\partial(X_3\beta_3)'} = 2M_1 \tag{2.18}$$

and is p.s.d.

Finally, the distance between the two estimated models H_1 and H_2 is defined, analogously to (2.7) and (2.12), as

$$\begin{aligned} D &= (\hat{\epsilon}_2 - \hat{\epsilon}_1)'(\hat{\epsilon}_2 - \hat{\epsilon}_1) \\ &= \hat{\epsilon}_2'\hat{\epsilon}_2 - 2\hat{\epsilon}_2'\hat{\epsilon}_1 + \hat{\epsilon}_1'\hat{\epsilon}_1. \end{aligned} \tag{2.19}$$

2.2. *A Model Selection Criterion*

The choice between H_1 and H_2 is made using model selection criteria. Several of these rules can be described as choosing the alternative hypothesis H_1 when

$$\hat{\epsilon}_2'\hat{\epsilon}_2 > q\hat{\epsilon}_1'\hat{\epsilon}_1 \qquad \text{with} \quad q > 1, \tag{2.20}$$

where q is a function of T, k_1, and k_2. It is easy to see that increasing q reduces the probability of choosing H_1. Adding the same terms to both sides gives

$$\hat{\epsilon}_2'\hat{\epsilon}_2 - 2\hat{\epsilon}_2'\hat{\epsilon}_1 + \hat{\epsilon}_1'\hat{\epsilon}_1 > q\hat{\epsilon}_1'\hat{\epsilon}_1 - 2\hat{\epsilon}_2'\hat{\epsilon}_1 + \hat{\epsilon}_1'\hat{\epsilon}_1. \tag{2.21}$$

Because $P_1P = P$, it can easily be verified that $M_1M = M$, and hence that $\hat{\epsilon}_1'\hat{\epsilon}_1 = \hat{\epsilon}_2'\hat{\epsilon}_1$.[3] Together with (2.19) this implies that a model selection criterion can be expressed as the choice for H_1 when the distance D exceeds a critical value Q, i.e.

$$D > Q, \qquad \text{where} \quad Q = (q-1)\hat{\epsilon}_1'\hat{\epsilon}_1. \tag{2.22}$$

[3]When $\hat{\epsilon}_1'\hat{\epsilon}_1 = \hat{\epsilon}_2'\hat{\epsilon}_1$, the distance D between models H_1 and H_2 [see (2.19)] can be written as $D = \hat{\epsilon}_2'\hat{\epsilon}_2 - \hat{\epsilon}_1'\hat{\epsilon}_1$. In Sneek [10, p. 105] the expectation of this expression is used for the distance between two estimated models. Analogously, D_i is defined as $\hat{\epsilon}_i'\hat{\epsilon}_i - \epsilon'\epsilon$, which is the distance between estimated model H_i and the data generating process, $i = 1,2$. It is however usually not true that $\hat{\epsilon}_i'\epsilon = \epsilon'\epsilon$, because this would imply that $X_3\beta_3 = 0$. In that case the true model would be (2.1), and this is not possible in our opinion. Hence, the definitions of D_i as in (2.7) and (2.12) are to be preferred.

With (2.3) it is possible to write D and Q in terms of $X_3\beta_3$ and ϵ:

$$\begin{aligned} D &= \hat{\epsilon}_2'\hat{\epsilon}_2 - \hat{\epsilon}_1'\hat{\epsilon}_1 \\ &= y'M_1y - y'My \\ &= (X_2\beta_2 + X_3\beta_3 + \epsilon)'(M_1 - M)(X_2\beta_2 + X_3\beta_3 + \epsilon) \qquad (2.23) \end{aligned}$$

and

$$Q = (q-1)\left[(X_3\beta_3 + \epsilon)'M(X_3\beta_3 + \epsilon)\right]. \qquad (2.24)$$

The scalars D and Q are symmetric parabolic functions of $X_3\beta_3$, and ≥ 0, because $M_1 - M = M_1P$ and M are p.s.d. matrices. The first derivatives of D and Q with respect to $X_3\beta_3$ are given by

$$\frac{\partial D}{\partial X_3\beta_3} = 2M_1P(X_2\beta_2 + X_3\beta_3 + \epsilon) \qquad (2.25)$$

and

$$\frac{\partial Q}{\partial X_3\beta_3} = 2(q-1)M(X_3\beta_3 + \epsilon). \qquad (2.26)$$

Assuming full rank conditions for X_2 and X_3, the minimal value 0 for D and Q will be reached when $X_3\beta_3 = -X_2\beta_2 - \epsilon$ and $X_3\beta_3 = -\epsilon$, respectively.

From the foregoing it follows clearly that $X_3\beta_3$ can have an effect on the model choice. It remains however difficult to determine its exact influence, because D_1, D_2, D, Q, and their respective first derivatives are not only affected by $X_3\beta_3$, but also by $X_2\beta_2$, M_1, M, and ϵ. Moreover, a large k_3 does not necessarily imply a substantial $X_3\beta_3$. It is possible that $X_3\beta_3$ is influential, while $k_3 = 1$. Conversely, it might be the case that, although k_3 is large, $X_3\beta_3 \approx 0$. Furthermore, it might occur that, although $X_3\beta_3$ is substantial, $MX_3\beta_3$ and $M_1X_3\beta_3$ are not. Anyway, the first derivatives in (2.25) and (2.26) will be equal with probability zero when (2.3) is the true model. Hence, when $X_3\beta_3$ changes (and thereby y), D and Q will change, but not in the same way.

Although in practice the true model and the impact of $X_3\beta_3$ are unknown, there is still something to say about the criteria. We might want to know how stable our choice is when we prefer e.g. H_1, for it can be the $X_3\beta_3$ that leads to this choice. Consider again (2.22), and suppose we choose H_1. If Q decreases in comparison with D because of $X_3\beta_3$, the probability of choosing H_1 $[P(H_1)]$ gets larger. A larger value for q "corrects" this influence of $X_3\beta_3$. Hence, when H_1 is chosen with the criterion with the largest q, then the influence of a possible true model on this choice is the least. Analogously, the true model has the smallest influence on the choice for H_2 if it is chosen with the criterion with the smallest q.

The above leads to the conclusion that when H_1 or H_2 is chosen with the criteria with the largest *and* the smallest q, the choice is less influenced by the correct specification. Nevertheless, this does not imply that we have approximated the true model best. In the case however when H_1 is preferred with some criteria, and H_2 with others, we should preferably postpone our choice, because we will never know if we made the "right" choice.

3. EIGHT MODEL SELECTION CRITERIA

In the first subsection we establish the factor q for eight model selection rules. In the second subsection these factors are compared.

3.1. *The Determination of the Factor q*

Consider first the F-test, i.e.

$$F=\frac{(\hat{\epsilon}_2'\hat{\epsilon}_2-\hat{\epsilon}_1'\hat{\epsilon}_1)/k_2}{\hat{\epsilon}_1'\hat{\epsilon}_1/(T-k_1-k_2)}. \tag{3.1}$$

The null hypothesis, H_2, is rejected when $F>c$, where c is the 95th percentile[4] of the F-distribution with k_2 and $T-k_1-k_2$ degrees of freedom. Rearranging this expression gives the F-test as

$$(T-k_1-k_2)\hat{\epsilon}_2'\hat{\epsilon}_2>(T-k_1-k_2+ck_2)\hat{\epsilon}_1'\hat{\epsilon}_1, \tag{3.2}$$

and the factor q is given by

$$\frac{T-k_1-k_2+ck_2}{T-k_1-k_2}=1+\frac{ck_2}{T-k_1-k_2}, \tag{3.3}$$

[4] The 95th percentile is used here because it is mostly used in practice. It could have been the 99th or 97.5th percentile, and that would not change the content of Table 2.

which exceeds 1 when $k_2 \neq 0$. Furthermore, it is interesting to notice that if the F-test can be expressed in the general form (2.20), all other criteria, to be discussed below, can be seen as F-tests with critical values other than c.

Another familiar model selection rule is the adjusted determination coefficient $\bar{R}^2$, which is

$$\bar{R}^2 = 1 - \left(\frac{T-1}{T-k}\right)\left(\frac{\hat{\epsilon}'\hat{\epsilon}}{y'y}\right). \tag{3.4}$$

Using $\bar{R}^2$ for the choice between (2.1) and (2.2) results in the selection of the first when

$$1 - \left(\frac{T-1}{T-k_1-k_2}\right)\left(\frac{\hat{\epsilon}_1'\hat{\epsilon}_1}{y'y}\right) > 1 - \left(\frac{T-1}{T-k_1}\right)\left(\frac{\hat{\epsilon}_2'\hat{\epsilon}_2}{y'y}\right)$$

$$\Leftrightarrow \left(\frac{T-1}{T-k_1}\right)\hat{\epsilon}_2'\hat{\epsilon}_2 > \left(\frac{T-1}{T-k_1-k_2}\right)\hat{\epsilon}_1'\hat{\epsilon}_1. \tag{3.5}$$

The coefficient q for $\bar{R}^2$ is given by

$$\frac{T-k_1}{T-k_1-k_2} = 1 + \frac{k_2}{T-k_1-k_2}. \tag{3.6}$$

It can easily be seen that the q in (3.3) exceeds the one in (3.6) because $c > 1$. Before comparing the criteria, we consider six other model selection rules, most of which are discussed in [4]. The notation used there is maintained here.

The Amemiya [2] prediction criterion (PC) induces the choice of model (2.1) when

$$\left(\frac{T+k_1}{T-k_1}\right)\hat{\epsilon}_2'\hat{\epsilon}_2 > \left(\frac{T+k_1+k_2}{T-k_1-k_2}\right)\hat{\epsilon}_1'\hat{\epsilon}_1. \tag{3.7}$$

The factor q of the PC criterion is

$$\frac{(T+k_1+k_2)(T-k_1)}{(T-k_1-k_2)(T+k_1)}. \tag{3.8}$$

This can be written as

$$1+\frac{2Tk_2}{(T-k_1-k_2)(T+k_1)}. \tag{3.9}$$

The fourth selection rule to be considered is Akaike's [1] information criterion (AIC), which prefers H_1 when

$$\exp\left(\frac{2k_1}{T}\right)\cdot\hat{\epsilon}_2'\hat{\epsilon}_2 > \exp\left(\frac{2(k_1+k_2)}{T}\right)\cdot\hat{\epsilon}_1'\hat{\epsilon}_1. \tag{3.10}$$

For this criterion the factor q is

$$\exp\left(\frac{2k_2}{T}\right). \tag{3.11}$$

It is interesting to note that it does not depend on k_1. Furthermore, it is easy to see that the factor in (3.11) is larger than 1 provided that $k_2 \neq 0$.

In Shibata [9] the fifth model selection rule can be found: The hypothesis (2.1) is preferred when

$$\left(1+\frac{2k_1}{T}\right)\hat{\epsilon}_2'\hat{\epsilon}_2 > \left(1+\frac{2k_1+2k_2}{T}\right)\hat{\epsilon}_1'\hat{\epsilon}_1. \tag{3.12}$$

For the Shibata criterion the q is

$$\frac{T+2k_1+2k_2}{T+2k_1}=1+\frac{2k_2}{T+2k_1}. \tag{3.13}$$

For estimating the dimension of the model, in Schwarz [8] a criterion is developed under which a model with many parameters will be heavily penalized (see [4]). It implies that the parsimonious model (2.2) is rejected only when

$$T^{k_1/T}\hat{\epsilon}_2'\hat{\epsilon}_2 > T^{(k_1+k_2)/T}\hat{\epsilon}_1'\hat{\epsilon}_1. \tag{3.14}$$

The Schwarz criterion implies a factor

$$T^{k_2/T}, \tag{3.15}$$

which does not depend on k_1 [cf. (3.11)]. After writing (3.15) in its Taylor series expansion

$$1+\frac{k_2(\log T)}{T}+\frac{1}{2}\frac{k_2^2(\log T)^2}{T^2}+\cdots, \tag{3.16}$$

it is clear that (3.15) exceeds 1.

The seventh model selection criterion is given in Craven and Wahba [3]. This rule has recently been appreciated as a selection rule for nested hypotheses (see [4]). This generalized cross validation (GCV) criterion rejects (2.2) if

$$\left(\frac{T}{T-k_1}\right)^2 \hat{\epsilon}_2'\hat{\epsilon}_2 > \left(\frac{T}{T-k_1-k_2}\right)^2 \hat{\epsilon}_1'\hat{\epsilon}_1. \tag{3.17}$$

In this case the factor q equals

$$\left(\frac{T-k_1}{T-k_1-k_2}\right)^2 = 1+\frac{2k_2}{(T-k_1-k_2)}+\frac{k_2^2}{(T-k_1-k_2)^2}. \tag{3.18}$$

The last criterion considered in this paper is the Rice [7] criterion. This one prefers H_1 if

$$\frac{1}{1-(2k_1/T)}\hat{\epsilon}_2'\hat{\epsilon}_2 > \frac{1}{1-[(2k_1+2k_2)/T]}\hat{\epsilon}_1'\hat{\epsilon}_1, \tag{3.19}$$

and its factor q is

$$\frac{T}{T-2k_1-2k_2}\Big/\frac{T}{T-2k_1} = 1+\frac{2k_2}{T-2k_1-2k_2}. \tag{3.20}$$

To facilitate the comparison in the next subsection, a summary of the above results is presented in Table 1.

3.2. *A Comparison of the Criteria*

The eight model selection criteria are compared with respect to their factors q. Denoting the latter as q_1 for the F-test through q_8 for the Rice

TABLE 1

EIGHT MODEL SELECTION CRITERIA AND THEIR CORRESPONDING FACTORS q

No.	Model selection criterion	Factor q_i
1	F-statistic	$1+\dfrac{ck_2}{T-k_1-k_2}$
2	$\bar{R}^2$	$1+\dfrac{k_2}{T-k_1-k_2}$
3	Amemiya's PC	$1+\dfrac{2Tk_2}{(T-k_1-k_2)(T+k_1)}$
4	Akaike's information criterion	$\exp\left(\dfrac{2k_2}{T}\right)$
5	Shibata	$1+\dfrac{2k_2}{T+2k_1}$
6	Schwarz	$T^{k_2/T}$
7	GCV (Craven and Wahba)	$1+\dfrac{2k_2}{T-k_1-k_2}+\dfrac{k_2^2}{(T-k_1-k_2)^2}$
8	Rice	$1+\dfrac{2k_2}{T-2k_1-2k_2}$

criterion (see Table 1), we establish if the inequality $q_i > q_j$ holds, where $i, j = 1,\dots,8$.

We make a distinction between three cases of combinations of k_1, k_2, and T. First, there are inequalities which will hold for every combination, provided that $k_1 \neq 0$, $k_2 \neq 0$, and $T > k_1, k_2$. An example is that it will always be true that $q_8 > q_2$, because

$$\frac{2k_2}{(T-2k_1-2k_2)} > \frac{k_2}{(T-k_1-k_2)} \Leftrightarrow 2T-2k_1-2k_2 > T-2k_1-2k_2$$
$$\Leftrightarrow T > 0. \tag{3.21}$$

Secondly, there are two inequalities valid under special conditions only, i.e., T is large and k_1, k_2 are relatively small (say $T/k_1 > 6$ and $T/k_2 > 6$). The first is $q_5 > q_2$, because

$$\frac{2k_2}{T+2k_1} > \frac{k_2}{T-k_1-k_2} \Leftrightarrow 2Tk_2 - 2k_1k_2 - 2k_2^2 - Tk_2 - 2k_1k_2 > 0 \tag{3.22}$$

is valid when $T > 4k_1 + 2k_2$ and $k_2 \neq 0$.

The second inequality applying under these conditions is $q_4 > q_2$. Taking the first two terms of the Taylor expansion of q_4 and writing $k_1 = aT$, $k_2 = bT$, with $0 < a, b < 1$, then $q_4 > q_2$ if

$$T^2(1 - 2a - 2ab - 2b^2) > 0 \tag{3.23}$$

If $a, b < \frac{1}{6}$, this will certainly be true.

Finally, the other inequalities will be called indefinite. An example is q_1 versus q_3:

$$\frac{ck_2}{T - k_1 - k_2} > \frac{2Tk_2}{(T - k_1 - k_2)(T + k_1)} \quad \Leftrightarrow \quad (c-2)\cdot T + ck_1 > 0. \tag{3.24}$$

When c corresponds with the 95th percentile, and k_1, k_2, and T are very large, (3.24) will not be true.

The investigation of the remaining 24 inequalities and of the conditions under which they are valid is not presented here in detail. The result is summarized in Table 2.

TABLE 2

THE MODEL SELECTION CRITERIA COMPARED[a]

i \ j	1	2	3	4	5	6	7	8
1		+	±	±	±	±	±	±
2			−	−[b]	−[b]	−	−	−
3				+	+	±	−	−
4					+	−	−	−
5						±	−	−
6							±	±
7								−
8								

[a] $+$: $q_i > q_j$ for all T, k_1, and k_2 ($k_1, k_2 \neq 0$ and $T > k_1, k_2$). $-$: $q_i < q_j$ for all T, k_1, and k_2 ($k_1, k_2 \neq 0$ and $T > k_1, k_2$). $\pm$: indefinite.

[b] The inequality is established for certain conditions of T, k_1, and k_2, i.e. $(T/k_1), (T/k_2) > 6$.

TABLE 3
q VALUES FOR $T = 100$, $k_1 = 2$, AND $k_2 = 1$

No.	Criterion	q
1	Schwarz	1.0471
2	F-test	1.0408
3	Rice	1.0213
4	GCV	1.0207
5	PC	1.0202
6	AIC	1.0202
7	Shibata	1.0192
8	$\bar{R}^2$	1.0103

4. A SMALL SIMULATION EXPERIMENT

Consider the hypotheses

$$H_1: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_1, \tag{4.1}$$

$$H_2: y = \beta_0 + \beta_1 x_1 + \epsilon_2, \tag{4.2}$$

where $k_1 = 2$ and $k_2 = 1$. The choice between H_1 and H_2 is made with criteria as in (2.20): choose for H_1 if

$$\tilde{q} > q, \quad \text{where} \quad \tilde{q} = \hat{\epsilon}_2'\hat{\epsilon}_2/\hat{\epsilon}_1'\hat{\epsilon}_1. \tag{4.3}$$

When $T = 100$, the values of q for the several criteria can be calculated (see Table 3).

Generate $x_1, \ldots, x_{25} \sim U[0,1]$, $\epsilon \sim N(0,1)$, and establish

$$y_k = 2 + \sum_{i=1}^{k} x_i + \epsilon \tag{4.4}$$

for $k = 2, \ldots, 25$. The experiment consists of the choice between H_1 and H_2 with (4.3), while y changes [because of (4.4)]. Although the models all have extremely low R^2-values, some results are salient enough to be worth displaying (Table 4). From this small experiment it can be seen that the true model (4.4) influences the choice between H_1 and H_2. If we use e.g. the GCV criterion alone (No. 4 in Tables 3 and 4), it depends on the y_k (cf. y_{21} and y_{22}) whether we choose H_1 or H_2. The problem however is that we do not

TABLE 4

THE RESULTS FROM A SMALL SIMULATION EXPERIMENT

y	$\tilde{q}$	Choose H_1 with criteria	Choose H_2 with criteria
y_5	1.1070	All	None
y_{10}	1.0503	All	None
y_{15}	1.0440	2, 3, 4, 5, 6, 7, 8	1
y_{20}	1.0275	3, 4, 5, 6, 7, 8	1, 2
y_{21}	1.0230	3, 4, 5, 6, 7, 8	1, 2
y_{22}	1.0162	8	1, 2, 3, 4, 5, 6, 7
y_{23}	1.0094	None	All

know what has generated y in practice. Finally, it should be noted that the eight criteria are, of course, not the only ones to be used for model selection. Diagnostic tests, specification tests, and all kinds of prior knowledge should also be used.

5. DISCUSSION AND CONCLUDING REMARKS

This paper deals with aspects of the testing of the nested standard linear hypotheses (2.1) versus (2.2). First, we assumed that there exists a "true" model (2.3). The models compared are related to this correct specification using a distance measure. When the true specification differs substantially from the "wrong" models, the respective distances D_1 and D_2 [see (2.7) and (2.12)] can likewise be different.

Secondly, a model selection criterion in its most general form (2.20) can be seen as the choice of specification (2.1) when the distance D between H_1 and H_2, after estimation, exceeds a certain critical value Q. In Section 2 we have demonstrated that this D and Q are influenced by the extent to which the true specification differs from the estimated wrong models. This affects the choice between (2.1) and (2.2). The factor q, which is a function of k_1, k_2, and T, is an important factor in this influence.

Finally, q will have different values for various model selection criteria. The values for eight choice rules are established in Section 3. It can be seen from Table 2 that in most cases $\bar{R}^2$ has a low value for the factor q, and that the Schwarz, the GCV, and the Rice criteria have high values.

In practice, however, we do not know the true specification, so we use model selection criteria as in (2.20). When the sum of squared estimated residuals of model H_2 exceeds a factor q times the corresponding sum of model H_1, the latter is preferred. The larger q, the smaller the probability that the largest model is chosen. Engle and Brown [4] find that the criteria with the largest q perform better in selecting forecasting models.

In this paper it is demonstrated that not only q, but also the hypothesized true specification, can influence the choice of the "best" model. This has been illustrated with a small simulation experiment. It confirms the simulation results in Sneek [10, p. 300] that the probability of selecting a given model varies with the distance between this model and the generating process. This has also an implication for the use of selection criteria. If we make a choice between H_1 and H_2 in practice, we only have to use two statistical selection criteria: the one with the smallest q and the one with the largest q. If with both criteria a certain hypothesis is preferred, a choice is made on which the "true" model has the least effect. If however these criteria prefer each another model, then it is very likely that the choice has been influenced by the correct, but unknown, specification in an untraceable way. In that case we can better construct other models between which we must choose.

REFERENCES

1 H. Akaike, A new look at the statistical model identification, *IEEE Trans. Automat. Control* AC-19:716–723 (1974).

2 T. Amemiya, Selection of regressors, *Internat. Econom. Rev.* 21:331–354 (1980).

3 P. Craven and G. Wahba, Smoothing noisy data with spline functions, *Numer. Math.* 31:377–403 (1979).

4 R. F. Engle and S. J. Brown, Model selection for forecasting, *Appl. Math. Comput.* 20:313–327 (1986).

5 J. F. Kiviet, *Testing Linear Econometric Models*, Univ. of Amsterdam, Amsterdam, 1987.

6 T. Kloek, Note on a large-sample result in specification analysis, *Econometrica* 43:933–936 (1975).

7 J. Rice, Bandwidth choice for nonparametric regression, *Ann. Statist.* 12:1215–1230 (1984).

8 G. Schwarz, Estimating the dimension of a model, *Ann. Statist.* 6:461–464 (1978).

9 R. Shibata, An optimal selection of regression variables, *Biometrika* 68:45–54 (1981).

10 J. M. Sneek, *Modelling Procedures for Univariate Economic Time Series*, Free University Press, Amsterdam, 1984.