The Annals of Mathematical Statistics
1970, Vol. 41, No. 3, 956–969

THE CONSISTENCY OF NONLINEAR REGRESSIONS[1]

BY E. MALINVAUD

1. Introduction. A sample of T observations on the variables x, z_1, $z_2, \cdots$, $z_m (j = 1, 2 \cdots m)$ has been generated according to the model:

$$x_t = g(z_{1t}, z_{2t} \cdots z_{mt}; \alpha_1, \alpha_2 \cdots \alpha_p) + \varepsilon_t, \tag{1}$$

in which x_t, z_{1t}, $z_{2t} \cdots z_{mt}$ designate the values taken by the variables in observation $t (t = 1, 2 \cdots T)$, $\alpha_1, \alpha_2 \cdots \alpha_p$ are p unknown parameters to be estimated $(k = 1, 2 \cdots p)$, ε_t is an unobservable random variable with zero expected value and g is a known function of its $m+p$ arguments. By a regression on model (1) we mean the computation of estimates $\hat{\alpha}_1, \hat{\alpha}_2 \cdots \hat{\alpha}_p$ that minimize the mean square deviation of x from g:

$$T^{-1} L_T(\alpha) = T^{-1} \sum_{t=1}^{T} [x_t - g(z_{1t} \cdots z_{mt}; \alpha_1 \cdots \alpha_p)]^2. \tag{2}$$

The regression is said to be linear if the function g is linear in the vector α with components $\alpha_1, \alpha_2 \cdots \alpha_p$, a situation extensively explored in the literature.

Cases of model (1) with nonlinear functions g occur in various areas of applied statistics. They are frequent in econometrics, where the model often results from rather specific theories implying special forms for the dependence between the variables z_{jt} considered as exogenous and the variable x_t taken to be endogenous.

An important particular case of model (1) occurs when the function g is linear with respect to the variables z_{jt}:

$$x_t = \sum_{j=1}^{m} a_j(\alpha_1 \cdots \alpha_p) \cdot z_{jt} + \varepsilon_t \tag{3}$$

a_j being known functions of the parameters. In econometrics, distributed lag models are of this type, whereas any overidentified simultaneous equation system has a reduced form corresponding to the multivariate generalization of model (3), a generalization about which I shall make some comments in the concluding section.

The regression method for estimating the α_k is of widespread use and has some a priori appeal.[2] As is well known, important properties of linear regressions do not depend on the particular form assumed by the distribution of the random term ε_t, a feature that has definite advantages in most fields of applied statistics and that extends to asymptotic properties of nonlinear regressions.

Such being the case one may be surprised to realize how little developed is the statistical theory of nonlinear regressions. Research has been concentrated on the

Received August 23, 1967; revised June 16, 1969.

[1] This paper was written in the spring of 1967 during a visit at Berkeley and was supported by the Office of Naval Research under Contract Nonr-222(77) with the University of California.

[2] I probably should mention here, even though I shall not refer to it again, that the regression method gives maximum likelihood estimates if the ε_t are normally, independently and identically distributed with zero expected value.

problems raised by the computation of the estimates minimizing $L_T(\alpha)$. But little effort has been spent in exploring the conditions under which nonlinear regressions perform well.

We cannot reasonably expect to build a small sample theory that would apply when g is not linear and the distribution of ε_t is not fully specified. On the other hand, an asymptotic theory seems to be well within reach. As is often asserted, the minimizing vector $\hat{\alpha}$ should have the same asymptotic stochastic properties as a vector $\tilde{\alpha}$ minimizing an expression similar to $L_T(\alpha)$, but in which $g(z_{1t}\cdots z_{mp}; \alpha_1\cdots\alpha_p)$ would be replaced by a linear approximation around the true value α^0 of α:

$$g(z_{1t}\cdots z_{mt}; \alpha_1{}^0\cdots\alpha_p{}^0)+\sum_{k=1}^{p}(\alpha_k-\alpha_k{}^0)\frac{\partial}{\partial\alpha_k}g(z_{1t}\cdots z_{mt}; \alpha_1{}^0\cdots\alpha_p{}^0). \tag{4}$$

The limit properties of such a vector $\tilde{\alpha}$ directly follow from linear regression theory. It is, however, clear that such an asymptotic equivalence between $\hat{\alpha}$ and $\tilde{\alpha}$ cannot be established unless one has first proved the consistency of $\hat{\alpha}$, i.e., the tendency of $\hat{\alpha}$ to α^0 as T increases indefinitely. This note is intended to explore the conditions under which consistency holds; it is thus concerned with the first and essential step in the statistical theory of nonlinear regressions.[3]

Before the general theory is approached, two examples in which consistency does not hold may be considered briefly; they will provide a background for the following discussion.

EXAMPLE 1. Consider the relation

$$x_t = e^{-\alpha t}+\varepsilon_t \tag{5}$$

where x_t is the observed value of x for period t (the periods being numbered $t = 1, 2\cdots T$), α is a positive number to be estimated, the ε_t are independently and identically distributed random variables having expected value zero and standard deviation σ.

The regression estimate $\hat{\alpha}$ satisfies the following equation:

$$\sum_{t=1}^{T} t e^{-\hat{\alpha}t}(x_t-e^{-\hat{\alpha}t}) = 0, \tag{6}$$

---

[3] When writing my book, E. Malinvaud (1966), I found it useful to base the presentation of various econometric methods on the theory of nonlinear regression. I therefore devoted one chapter to this theory (Chapter 9). The present paper studies more deeply the question of consistency. On the other hand, the implications of consistency for other asymptotic properties of $\hat{\alpha}$ may be found in the above reference.

When this paper was in its final stage of completion, R. I. Jenrich (1969) dealt with the same subject, assuming throughout a compact Ω. His Theorem 6 is similar to my Theorem 2 and actually proves *strong* consistency of $\hat{\alpha}$. On the other hand, he does not consider how my Assumption 4 could be derived from more basic hypotheses about the function g and the sequence of the exogenous variables z_t. The reader may also prefer my elementary but tedious proof of Theorem 2 to his more elegant proof that uses a number of non-classical mathematical properties.

which, in view of (5), may also be written as: $u_T+v_T+w_T=0$ with:

$$u_T=\sum_{t=1}^{T} t\,e^{-\alpha^0 t}\varepsilon_t$$
$$v_T=\sum_{t=1}^{T} t\,e^{-\hat{\alpha}t}(e^{-\alpha^0 t}-e^{-\hat{\alpha}t})$$
$$w_T=\sum_{t=1}^{T} t(e^{-\hat{\alpha}t}-e^{-\alpha^0 t})\varepsilon_t.$$

Let us suppose that $\hat{\alpha}$ is consistent, i.e., $\hat{\alpha}$ tends in probability to the true value α^0 when T increases indefinitely. Then, if $\alpha^0>0$, it can be proved that u_T has a limit distribution with a positive variance, whereas v_T and w_T tend in probability to zero. This contradiction proves that $\hat{\alpha}$ cannot be consistent.

This example is similar to the following one from linear regression theory: $x_t=\alpha/t+\varepsilon_t$, $t=1, 2\cdots T$, in which the variance of $\hat{\alpha}$ tends to a positive limit, because $\sum_t 1/t^2$ does not increase indefinitely with T.

EXAMPLE 2. Let α be in the real unit interval closed on the left, open on the right, and $g(t;\alpha)$ be the tth digit in the binary expansion of α. Let the ε_t be independently and identically distributed with the (known) binomial distribution giving probabilities 3/8 and 5/8 respectively to the values $-4/3$ and 4/5.

We immediately see that x_t can take only four values: $-4/3$ and 4/5 corresponding to $g(t;\alpha)=0$, $-1/3$ and 9/5 corresponding to $g(t;\alpha)=1$. Hence α is identifiable, its tth binary digit being unambiguously determined by the tth observation. However, a regression will associate the digit 1 with 4/5 and 9/5, the digit 0 with $-4/3$ and $-1/3$. The resulting estimate $\hat{\alpha}$ will, of course, be quite different from α^0.

2. A Lemma. In model (1) it is assumed that the variables $z_{1t}\cdots z_{mt}$ are non-random. The following compact notation will be convenient:

$$g_t(\alpha)=g(z_{1t}, z_{2t}\cdots z_{mt};\alpha_1,\alpha_2\cdots\alpha_p). \tag{7}$$

Moreover, suppose that the vector α of the parameters may be restricted a priori to a proper subset of R^p, which we shall designate by Ω. The regression estimate $\hat{\alpha}$ will minimize in Ω the quantity:

$$L_T(\alpha)=\sum_{t=1}^{T}[x_t-g_t(\alpha)]^2. \tag{8}$$

If either g_t is not continuous or Ω is not compact, $\hat{\alpha}$ may not exist. When it exists it may not be unique. We shall not be concerned here with these questions, but shall consider the properties of one $\hat{\alpha}$ minimizing (8).

We begin with a general lemma that provides a criterion for consistency. While the meaning of this condition is not transparent, the lemma provides the foundation for deriving more useful results.

Let us introduce some notation:

$$q_t(\alpha)=g_t(\alpha)-g_t(\alpha^0), \tag{9}$$

where α^0 is the true value of the vector α;

$$Q_T(\alpha)=\sum_{t=1}^{T} q_t^2(\alpha). \tag{10}$$

When $Q_T(\alpha)$ is positive, we can also define:

$$\lambda_{tT}(\alpha) = q_t(\alpha)/Q_T(\alpha), \tag{11}$$

$$u_T(\alpha) = \sum_{t=1}^{T} \lambda_{tT}(\alpha)\varepsilon_t. \tag{12}$$

When $Q_T(\alpha)$ is zero we shall conventionally define $u_T(\alpha)$ as being zero. We shall prove:

LEMMA. *Suppose that, for every closed set ω not containing α^0,*
(i) $\inf_{\alpha \in \omega} Q_T(\alpha) > 0$ *for T sufficiently large;*
(ii) $\Pr\{\sup_{\alpha \in \omega} u_T(\alpha) \geqq \frac{1}{2}\}$ *tends to zero as T increases indefinitely. Then $\hat{\alpha}$ is a consistent estimate of α^0.*

From the definition of $\hat{\alpha}$, $Q_T(\alpha)$ and $u_T(\alpha)$ it follows that:

$$\sum_{t=1}^{T} \varepsilon_t^2 = \sum_{t=1}^{T} [x_t - g_t(\alpha^0)]^2 \geqq \sum_{t=1}^{T} [x_t - g_t(\hat{\alpha})]^2 = \sum_{t=1}^{T} \varepsilon_t^2 + Q_T(\hat{\alpha})[1 - 2u_T(\hat{\alpha})].$$

Hence:

$$Q_T(\hat{\alpha})[2u_T(\hat{\alpha}) - 1] \geqq 0; \tag{13}$$

either $Q_T(\hat{\alpha}) = 0$, or $u_T(\hat{\alpha}) \geqq \frac{1}{2}$.

Let ω be a closed set not containing α^0 and suppose $\hat{\alpha} \in \omega$; then either $\inf_{\alpha \in \omega} Q_T(\alpha) = 0$ or $\sup_{\alpha \in \omega} u_T(\alpha) \geqq \frac{1}{2}$. But the first possibility is excluded by (i) for T sufficiently large. Hence:

$$\Pr\{\hat{\alpha} \in \omega\} \leqq \Pr\{\sup_{\alpha \in \omega} u_T(\alpha) \geqq \tfrac{1}{2}\},$$

and the conclusion follows.[4]

The conditions of the lemma simultaneously involve properties of the random disturbances ε_t, of the explanatory variables z_{jt}, and of the function g. Their significance with respect to those basic elements of the model is obscure, and we must look for general cases in which the conditions may be proved to hold. We begin with the simple situation of equation (3) in which g is linear with respect to the explanatory variables z_{jt}.

3. The constrained linear model. Let $g_t(\alpha)$ have the form:

$$g_t(\alpha) = \sum_{j=1}^{m} a_j(\alpha) z_{jt}. \tag{14}$$

We designate by $a(\alpha)$ and z_t the m-vectors with components $a_j(\alpha)$ and z_{jt}. The model may then be written as:

$$x_t = a(\alpha)' \cdot z_t + \varepsilon_t, \tag{15}$$

where $a(\alpha)'$ is the transpose of the column vector $a(\alpha)$.

The fact that α belongs to Ω implies that $a(\alpha)$ belongs to the set $A = a(\Omega)$ in R^m. The interesting case arises when A is a proper subset of R^m. The model (15) may

[4] This simple proof of the lemma was suggested by a referee.

then be viewed as a linear regression model in which the coefficients of the variables are subject to some constraints.

The determination of the regression estimate $\hat{\alpha}$ may be formally decomposed into two steps:

(i) find $\hat{a}$ in A minimizing:[5]

$$L_T(a) = \sum_{t=1}^{T} (x_t - a' z_t)^2, \tag{16}$$

(ii) find $\hat{\alpha}$ in Ω such that $a(\hat{\alpha}) = \hat{a}$.

The proof of consistency may similarly be decomposed into two steps, one concerning the convergence of $\hat{a}$ to $a^0 = a(\alpha^0)$, the other the convergence of $\hat{\alpha}$ to α^0.

As is clear from the condition of the lemma, the fact that a is restricted to a subset A of R^m does not fundamentally complicate the proof of the convergence of $\hat{a}$ to a^0. The conditions that guarantee the consistency of regression estimates in the unconstrained linear model should also guarantee the consistency of $\hat{a}$ as an estimate of a^0 in the constrained model.

However, the convergence of $\hat{\alpha}$ to α^0 obviously requires a condition on the vector function $a(\alpha)$. The true value α^0 must be identified as the unique solution of the equation $a(\alpha) = a^0$. Moreover, convergence of $\hat{a}$ to a^0 must imply the convergence to α^0 of any solution of $a(\alpha) = \hat{a}$.

It is not necessary to look here for the greatest generality about the convergence of $\hat{a}$ to a^0. We shall limit ourselves to the two following conventional assumptions:

ASSUMPTION 1. The disturbances ε_t are independently and identically distributed with expected value zero and finite variance σ^2.

ASSUMPTION 2. The second-order moment matrix:

$$M_{zz} = T^{-1} \sum_{t=1}^{T} z_t z_t' \tag{17}$$

tends to a nonsingular matrix $\bar{M}$ when T increases indefinitely.

On the m functions $a_j(\alpha)$ we shall make the following hypothesis:

ASSUMPTION 3. Given any sequence $\{\alpha^T\}$ of vectors in Ω (for $T = 1, 2 \cdots$ ad infinitum), if $a(\alpha^T)$ converges to $a(\alpha^0)$ in R^m then α^T itself converges to α^0. (The inverse mapping a^{-1} from $a(\Omega) \subset R^m$ to Ω is one-to-one and continuous at $a(\alpha^0)$.)

Examples in which Assumption 3 does not hold are easily found. With $p = m = 1$, the functions $a(\alpha) = \alpha^2 - \alpha$ and $a(\alpha) = \alpha e^{-\alpha}$ do not fulfill it when $\alpha^0 = 0$; the first one because $\alpha = 1$ gives $a(1) = 0 = a(\alpha^0)$, the second one because $a(\alpha^T)$ tends to $a(\alpha^0) = 0$ when α^T increases indefinitely.

We can now easily prove:

THEOREM 1. *If Assumptions* 1 *and* 2 *hold, the constrained linear regression estimate* $\hat{a}$ *minimizing* (16) *is a consistent estimate of* a^0. *If, in addition, Assumption* 3 *holds, the corresponding* $\hat{\alpha}$ *is a consistent estimate of* α^0.

---

[5] In the absence of further assumptions, $\hat{a}$ may not exist. This is the place to point out that we are not concerned in this paper with the problem of existence of regression estimates.

The second assertion of the theorem follows directly from the first one and from the fact that a^{-1} is continuous. The first assertion will be seen to follow from the more general Theorem 2. However, a direct and much simpler proof will also be given here for a better understanding of the property.

Going back to the notation introduced at the beginning of Section 2, we may write:

$$Q_T(a) = T(a-a^0)'M_{zz}(a-a^0), \tag{18}$$

$$\lambda_{tT}(a) = (a-a^0)'z_t/Q_T(a), \tag{19}$$

$$Q_T(a)\cdot|u_T(a)| = \left|\sum_{t=1}^T (a-a^0)'z_t\varepsilon_t\right| \leqq T\sum_{j=1}^m |a_j-a_j^0|\cdot\left|T^{-1}\sum_{t=1}^T z_{jt}\varepsilon_t\right|. \tag{20}$$

Assumptions 1 and 2 imply that $T^{-1}\sum_{t=1}^T z_{jt}\varepsilon_t$ tends in the mean square to zero, hence also in probability.

Let C be any closed subset of A not containing a^0. We shall prove that $Q_T(a) > 0$ for all $a \in A$ and T sufficiently large and that

$$\sup_{a\in C} T|a_j-a_j^0|/Q_T(a) \tag{21}$$

is bounded by a quantity which does not depend on T. This and (20) imply that $\sup_{a\in C}|u_T(a)|$ tends in probability to zero; i.e., that both conditions of the lemma are fulfilled for $\hat{a}$.

Let $d(a)$ be the distance from a to a^0: $d^2(a) = \sum_{j=1}^m (a_j-a_j^0)^2$. Then we know that:

$$|a_j-a_j^0| \leqq d(a), \tag{22}$$

$$T^{-1}Q_T(a) \geqq v_T\, d^2(a), \tag{23}$$

where v_T is the smallest characteristic root of M_{zz} (see for instance E. Malinvaud (1966) footnote page 292). The convergence of M_{zz} to $\bar{M}$ implies the convergence of v_T to the smallest root $\bar{v}$ of $\bar{M}$. Moreover, $\bar{v}$ is positive because $\bar{M}$ is positive definite. Hence, at least for T sufficiently large, $v_T > \frac{1}{2}\bar{v}$ and:

$$T|a_j-a_j^0|/Q_T(a) < 2/\bar{v}\, d(a). \tag{24}$$

Since C is closed and does not contain a^0, $d(a)$ is bounded below in C by a positive number. Hence $Q_T(a) > 0$ for all $a \in A$ and T sufficiently large, and (21) is finite, which completes the proof of Theorem 1.

The constrained linear model covers many cases of nonlinear regression, often after a redefinition of explanatory variables and parameters. In particular some of the components of α may take only integer values or be dichotomous (α_k being either 0 or 1). Theorem 1 then gives an asymptotic justification to the procedure which consists in choosing the values of α so as to maximize the traditional squared multiple correlation coefficient R^2.

As an example consider the following distributed lag model discussed by I. Fisher (1937):

$$x_t = \sum_{\tau=0}^{\alpha_2-1} \alpha_1(1-\tau/\alpha_2)z_{t-\tau} + \varepsilon_t, \tag{25}$$

$t = 1, 2 \cdots T$ referring to observations ordered in time and $z_{t-\tau}$ designating the value of z_t lagged by τ periods. Suppose that α_1 may be any real number and α_2 any positive integer smaller or equal to m : $\alpha_2 \in \{1, 2 \cdots m\}$. Then define:

$$
\begin{aligned}
z_{jt} &= z_{t-j+1} && j = 1, 2 \cdots m \\
a_j(\alpha_1, \alpha_2) &= \alpha_1(1-(j-1)/\alpha_2) && j = 1, 2 \cdots \alpha_2 \\
&= 0 && j = \alpha_2+1 \cdots m.
\end{aligned} \tag{26}
$$

The regression estimate may be determined as follows:

(i) for any fixed value of α_2 ($\alpha_2 = 1, 2 \cdots m$), compute the homogenous regression of x_t with respect to $\sum_{\tau=0}^{\alpha_2-1}(1-\tau/\alpha_2)z_{t-\tau}$ and compute the resulting R^2;

(ii) choose for $\hat{\alpha}_2$ the value of α_2 that leads to the highest value for R^2.

The inverse function a^{-1} referred to in Assumption 3 may here be determined as giving: $\alpha_1 = a_1$, $\alpha_2 = a_1/(a_1-a_2)$. It is continuous at a^0 if ${a_1}^0 \neq {a_2}^0$, which is always true if and only if ${\alpha_1}^0 \neq 0$.

4. A first general result. The proof of Theorem 1 makes direct use of the fact that $u_T(\alpha)$ is the sum of a finite number of terms, each of which is the product of a nonstochastic function of α (through a) and a random variable independent of α and converging to zero with T: see relation (20) above. Such a decomposition cannot be applied when $g_t(\alpha)$ is not linear in z_t. We must therefore look for a different strategy of proof in the general case.

The following one will be explored now. Given any closed set ω not containing α^0, can we partition it into a finite number of subsets ω^2 (with $r = 1, 2 \cdots s, s+1$) in such a way that the conditions of the lemma hold for each ω^r? If we can do so with a partition independent of T, then the conditions of the lemma will hold for ω also. This can easily be seen. For instance with the second condition we can write:

$$
\Pr\{\sup_{\alpha \in \omega} u_T(\alpha) \geqq \tfrac{1}{2}\} \leqq \sum_{r=1}^{s+1} \Pr\{\sup_{\alpha \in \omega^r} u_T(\alpha) \geqq \tfrac{1}{2}\} \tag{27}
$$

because $u_T(\alpha)$ reaches $\frac{1}{2}$ in ω if and only if it reaches it in at least one of the ω^r. If each term of the sum of the right-hand member tends to zero, then the left-hand member must tend to zero.

For any two α and β contained in Ω, define the function $\varphi_T(\alpha, \beta)$ by:

$$
\varphi_T(\alpha, \beta) = T^{-1}\sum_{t=1}^{T}[g_t(\alpha)-g_t(\beta)]^2. \tag{28}
$$

We observe that $Q_T(\alpha)$ defined by (10) is equal to $T\varphi_T(\alpha, \alpha^0)$.

To Assumption 1 let us now add:

ASSUMPTION 4. There is a positive number ∂, a compact subset K of Ω containing α^0, and a number T_0 such that:

(i) the following inequality holds for all α not in K and all T larger than T_0:

$$
T^{-1}Q_T(\alpha) \geqq 4\sigma^2 + \partial; \tag{29}
$$

(ii) as T increases indefinitely the function $\varphi_T(\alpha, \beta)$ tends uniformly on $K \times K$

to a continuous function $\varphi(\alpha, \beta)$; the function of α defined by $\varphi(\alpha, \alpha^0)$ is zero only for $\alpha = \alpha^0$.

We shall comment later on this assumption, but immediately prove:

THEOREM 2. *If Assumptions* 1 *and* 4 *hold, then* $\hat{\alpha}$ *is a consistent estimate of* α^0.

Given any closed set ω of Ω not containing α^0, we must partition ω into $s+1$ subsets ω^r in such a way that the conditions of the lemma are satisfied for each of these subsets.

Let us first define ω^{s+1} to be the intersection of ω with the complement of K in Ω. Assumption 4 implies that the first condition of the lemma holds for ω^{s+1}. Applying the Schwarz inequality to the definition of $u_T(\alpha)$ as given in (10)–(12), we find:

$$|u_T(\alpha)|^2 \leqq \sum_{t=1}^T \varepsilon_t^2 \cdot \sum_{t=1}^T \lambda_{tT}^2(\alpha) = T^{-1} \sum_{t=1}^T \varepsilon_t^2 / [T^{-1} Q_T(\alpha)]. \tag{30}$$

As T increases, the numerator of the right-hand side converges with probability 1 to σ^2 in view of Assumption 1. For $T > T_0$ the denominator exceeds $4\sigma^2 + \partial$ as a consequence of (i) in Assumption 4. Hence the probability that there exist a point α in ω^{s+1} such that the fraction is at least equal to $\frac{1}{4}$ tends to zero with T, as is required by the second condition of the lemma.

It remains now to partition $\omega \cap K$ into a finite number s of subsets ω^r in such a way that the conditions be fulfilled for each of them. The set $\omega \cap K$ is compact, and does not contain α^0; Assumption 4 then implies the existence of two positive numbers $\underline{S}^2$ and $\bar{S}^2$ such that:

$$\underline{S}^2 < \phi(\alpha, \alpha^0) < \bar{S}^2 \quad \text{for all} \quad \alpha \in \omega \cap K. \tag{31}$$

Moreover, the continuity of $\phi(\alpha, \beta)$ and of $1/\phi(\alpha, \alpha^0)$, considered as functions of α on the compact set $\omega \cap K$, implies that we can find a finite partition of $\omega \cap K$ into sets ω^r, and a vector α^r in each ω^r, in such a way that:

$$\phi(\alpha, \alpha^r) \leqq \underline{S}^4/(100\sigma^2) \quad \text{for all} \quad \alpha \in \omega^r \quad \text{and all} \quad r, \tag{32}$$

$$\left|[\phi(\alpha, \alpha^0)]^{-1} - [\phi(\alpha^r, \alpha^0)]^{-1}\right| \leqq 1/(10\sigma\bar{S}) \quad \text{for all} \quad \alpha \in \omega^r \quad \text{and all} \quad r. \tag{33}$$

We shall see that this implies the second condition of the lemma for each ω^r, whereas (31) obviously implies the first condition.

Indeed, consider:

$$\sup_{\alpha \in \omega^r} u_T(\alpha) \leqq \sup_{\alpha \in \omega^r} [u_T(\alpha) - u_T(\alpha^r)] + u_T(\alpha^r). \tag{34}$$

The last term tends in probability to zero because:

$$E[u_T^2(\alpha^r)] = \sigma^2/Q_T(\alpha^r) = \sigma^2/[T\varphi_T(\alpha^r, \alpha^0)] \tag{35}$$

tends to zero as $\varphi_T(\alpha^r, \alpha^0)$ tends to $\varphi(\alpha^r, \alpha^0) \neq 0$. The second condition of the lemma will therefore be satisfied if

$$\Pr\{\sup_{\alpha \in \omega^r} [u_T(\alpha) - u_T(\alpha^r)] \geqq \tfrac{1}{3}\} \tag{36}$$

tends to zero with T.

The equality (12) defining $u_T(\alpha)$ and the Schwarz inequality imply that

$$[u_T(\alpha)-u_T(\alpha^r)]^2 \leqq T^{-1}\sum_{t=1}^{T}\varepsilon_t^2 \cdot T\sum_{t=1}^{T}[\lambda_{tT}(\alpha)-\lambda_{tT}(\alpha^r)]^2. \tag{37}$$

The probability that $T^{-1}\sum\varepsilon_t^2$ exceeds $16\sigma^2/9$ tends to zero with T. Hence the probability (36) will also tend to zero if, for all T sufficiently large,

$$\sup_{\alpha\in\omega^r}\{T\sum_{t=1}^{T}[\lambda_{tT}(\alpha)-\lambda_{tT}(\alpha^r)]^2\} \leqq 1/(16\sigma^2), \tag{38}$$

a property that we are now going to prove.

By the definition of $\lambda_{tT}(\alpha)$ in (11), we can write

$$\lambda_{tT}(\alpha)-\lambda_{tT}(\alpha^r) = \frac{q_t(\alpha)-q_t(\alpha^r)}{Q_T(\alpha^r)} + q_t(\alpha)\left[\frac{1}{Q_T(\alpha)}-\frac{1}{Q_T(\alpha^r)}\right]. \tag{39}$$

The triangular inequality in R^T, together with equations (10), (9) and (28), imply that

$$\left[\sum_{t=1}^{T}[\lambda_{tT}(\alpha)-\lambda_{tT}(\alpha^r)]^2\right]^{\frac{1}{2}} \leqq \frac{1}{Q_T(\alpha^r)}[T\varphi_T(\alpha,\alpha^r)]^{\frac{1}{2}} + \left|\frac{1}{Q_T(\alpha)}-\frac{1}{Q_T(\alpha^r)}\right| Q_T^{\frac{1}{2}}(\alpha). \tag{40}$$

To establish that (38) holds for all T sufficiently large, we need only prove the two following inequalities:

$$[\varphi_T(\alpha,\alpha^r)]^{\frac{1}{2}} \leqq \frac{1}{8\sigma}\cdot\frac{1}{T}Q_T(\alpha^r) \quad \text{for all} \quad \alpha\in\omega^r \quad \text{and all large} \quad T. \tag{41}$$

$$\left|\frac{T}{Q_T(\alpha)}-\frac{T}{Q_T(\alpha^r)}\right| \leqq \frac{1}{8\sigma}\cdot\left[\frac{1}{T}Q_T(\alpha)\right]^{-\frac{1}{2}} \quad \text{for all} \quad \alpha\in\omega^r \quad \text{and all large} \quad T. \tag{42}$$

Let us first consider (41). We can write the inequality as:

$$\varphi_T(\alpha,\alpha^r) \leqq \frac{1}{64\sigma^2}\varphi_T{}^2(\alpha^r,\alpha^0). \tag{43}$$

The uniform convergence of φ_T to φ, together with inequalities (31) and (32) imply that, for T sufficiently large and all $\alpha\in\omega^r$:

$$\varphi_T(\alpha,\alpha^r) \leqq \frac{S^4}{80\sigma^2} \leqq \frac{1}{80\sigma^2}\varphi^2(\alpha^r,\alpha^0) \leqq \frac{1}{64\sigma^2}\varphi_T{}^2(\alpha^r,\alpha^0). \tag{44}$$

The inequality (43) or (41) is therefore proved for all $\alpha\in\omega^r$ and all T sufficiently large.

Let us now consider (42). The inequality can be written

$$\left|\frac{1}{\varphi_T(\alpha,\alpha^0)}-\frac{1}{\varphi_T(\alpha^r,\alpha^0)}\right| \leqq \frac{1}{8\sigma}[\varphi_T(\alpha,\alpha^0)]^{-\frac{1}{2}}. \tag{45}$$

The uniform convergence of $\varphi_T{}^{-1}$ to φ^{-1} in ω^r, together with the inequalities (31) and (33) imply that

$$\left|\frac{1}{\varphi_T(\alpha,\alpha^0)}-\frac{1}{\varphi_T(\alpha^r,\alpha^0)}\right| \leqq \frac{1}{9\sigma\bar{S}} \leqq \frac{[\varphi(\alpha,\alpha^0)]^{-\frac{1}{2}}}{9\sigma} \leqq \frac{1}{8\sigma}[\varphi_T(\alpha,\alpha^0)]^{-\frac{1}{2}}. \tag{46}$$

The inequality (45) or (42) holds therefore for all $\alpha \in \omega^r$ and all T sufficiently large. This completes the proof of Theorem 2.

Let us now examine Assumption 4, which was used in the proof of this theorem. We easily see that it is not satisfied in the two examples of inconsistency given at the beginning of this note.

In Example 1, $\varphi_T(\alpha, \beta)$ tends to zero with T for all (α, β) (the convergence is uniform only in subsets of Ω that are bounded away from zero). Indeed, if for instance $\alpha \leqq \beta$:

$$\varphi_T(\alpha, \beta) = T^{-1} \sum_{t=1}^{T} (e^{-\alpha t} - e^{-\beta t})^2 = T^{-1} \sum_{t=1}^{T} e^{-2\alpha t} [1 - e^{(\alpha-\beta)t}]^2$$
$$\leqq T^{-1} \sum_{t=1}^{T} e^{-2\alpha t} \leqq T^{-1} e^{-2\alpha}/(1 - e^{-2\alpha}).$$

The nonnegative sequence $\varphi_T(\alpha, \beta)$ is bounded above by a sequence tending to zero with T. It therefore tends to zero. Neither condition (i) nor condition (ii) of Assumption 4 is satisfied.

In Example 2, $\varphi_T(\alpha, \beta)$ will not tend to a limit for all (α, β). Even if we restrict Ω to a dense subset for which convergence holds, the limit $\varphi(\alpha, \beta)$ will not be continuous. If, for example, $\alpha^0 = \frac{1}{2}$, there exists an increasing sequence α^n converging to α^0 from below ($n = 1, 2 \cdots$ ad infinitum) and such that $\varphi(\alpha^n, \alpha^0)$ tend to 1 with n, whereas $\varphi(\alpha^0, \alpha^0) = 0$.

However,[6] Assumption 4 does hold in the unconstrained linear model $x_t = \alpha' z_t + \varepsilon_t$ satisfying Assumptions 1 and 2. Indeed $\varphi_T(\alpha, \beta) = (\alpha - \beta)' M_{zz}(\alpha - \beta)$; $\varphi_T(\alpha, \alpha^0)$ will exceed $4\sigma^2 + \partial$ outside of an appropriate compact domain K, for all T sufficiently large. The function $\varphi_T(\alpha, \beta)$ tends to the continuous function $(\alpha - \beta)' \overline{M}(\alpha - \beta)$ that is zero only for $\alpha = \beta$, and the convergence is uniform in $K \times K$. Thus Theorem 1 is a consequence of Theorem 2.

5. A second general result. Assumption 4 does not involve the random disturbance ε_t; but it is not yet directly expressed in terms of the basic elements of the model: the function g, the set Ω and the sequence of the z_t. We must now look for more

---

[6] It may be worth recording here an example that does not satisfy Assumption 4 and for which I was not able to prove either consistency of the regression estimate or its inconsistency. Let $x_t = \sin \alpha t + \varepsilon_t$, where Ω is the open interval $(0, 2\pi)$. If Assumption 1 holds for ε_t, the parameter α is identifiable in Ω. One easily sees that $\varphi_T(\alpha, \beta)$ tends to zero if $\alpha = \beta$, to 2 if $\alpha + \beta = 2\pi$ and $\alpha \neq \pi$, to $\frac{1}{2}$ if $\alpha = \pi$, $\beta \neq \pi$ or $\beta = \pi$, $\alpha \neq \beta$, and to 1 in all other cases. Hence $\varphi(\alpha, \beta)$ is not continuous and the convergence is not uniform at the points of discontinuity.

It seems intuitively likely that the regression estimate $\hat{\alpha}$ is consistent. However, a direct study of the problem leads to the condition that $\Pr\{\sup_{\alpha \in \omega} |w_T(\alpha)| \geqq \frac{1}{4}\}$ should tend to zero for all closed $\omega \subset \Omega$ not containing α^0, where $w_T(\alpha) = T^{-1} \sum_{t=1}^{T} \varepsilon_t \sin \alpha t$. The random function $w_T(\alpha)$ tends to zero for all α; but this does not imply that the sup of $w_T(\alpha)$ in a given domain also tends to zero unless one a priori limits Ω to a finite number of points in the interval $(0, 2\pi)$.

Note added in proof. It seems that consistency of $\hat{\alpha}$ in this model can be proved with the technique used by A. M. Walker in "On the estimation of a harmonic component in a time series with stationary residuals. I. Independent variables", The Manchester–Sheffield School of Probability and Statistics, September 1969.

basic assumptions from which Assumption 4 could be derived. The ones listed below are rather strong, and could perhaps be somewhat relaxed. They exhibit, however, a general class of problems in which consistency of nonlinear regression will hold.

The main difficulty seems to arise from the choice of the hypotheses to be made on the sequence of the z_t. Unless some rather specific conditions are imposed on the function g, it seems necessary to require that, as T increases, all vectors z_t remain within a compact set Z and that their distribution exhibits some stability. In order to formalize the latter requirement, we shall first associate with each finite sequence a measure μ_T on R^m giving the proportion of the T first vectors z_t that belong to any Borel set; we shall then assume that this measure weakly converges to a limit μ. Whereas μ_T has by definition a finite support for each T, the limit measure μ may very well be diffuse over Z.

For any Borel set V of R^m let $\mu_T(V)$ be equal to $1/T$ multiplied by the number of vectors z_t that are contained in V among the T first ones. The weak convergence of μ_T to a measure μ means that, for any bounded continuous real function $f(z)$, the integral $\int f\, d\mu_T$, which is the average value of f in the sample, tends to $\int f d\mu$ as T increases indefinitely.

ASSUMPTION 5. The vectors z_t are all contained in a compact set Z of R^m. The measure μ_T weakly converges to a measure μ.

We shall restrict attention to the case in which the function $g(z;\alpha)$ is continuous with respect to all its $m+p$ arguments z_j and α_k simultaneously. Hence:

ASSUMPTION 6. The function $g(z;\alpha)$ is continuous on $Z \times \Omega$.

We still need to remove explicitly a situation that corresponds to multicollinearity in linear regression, namely the situation in which the z_t would be such as to make two distinct values of α (asymptotically) undistinguishable by observation of the sample.

ASSUMPTION 7. The sequence $\{z_t\}$ separates Ω in the sense that, given any two distinct vectors α and β in Ω, the set of all z such that $g(z;\alpha) \neq g(z;\beta)$ has a positive measure μ.

Finally we must add a condition to the effect of checking (i) in Assumption 4, namely that, outside of a compact set K and for sufficiently large T, $T^{-1}Q_T(\alpha)$ be bounded below by a number larger than $4\sigma^2$. We could assume this condition explicitly, but it may appear unsatisfactory because it depends on the unknowns α^0 and σ^2. We shall therefore make an assumption that is sufficient for our purpose but obviously not necessary. We shall require that $g(z;\alpha)$ increase indefinitely with α in the following sense:

ASSUMPTION 8. There is some T_0 such that for any $T > T_0$ and any positive number G, the set of all α such that

$$T^{-1}\sum_{t=1}^{T} g^2(z_t;\alpha) \leqq G \tag{47}$$

is bounded; furthermore, this bound is uniform in T.

We can now state:

THEOREM 3. *If Assumptions* 1, 5, 6, 7 *and* 8 *hold, then* $\hat{\alpha}$ *is a consistent estimate of* α^0.

We must prove that Assumptions 5 to 8 imply Assumption 4. Let us first check that for any positive number ∂ we can find a compact set K so that (i) in Assumption 4 holds with the number T_0 specified in Assumption 8. Suppose such were not the case; there would then exist an unbounded sequence of vectors α^n (with $n = 1, 2 \cdots$ ad infinitum) such that

$$T^{-1} Q_T(\alpha^n) < 4\sigma^2 + \partial \quad \text{for arbitrarily large} \quad T. \tag{48}$$

Now, apply the triangular inequality in R^T to: $g(z_t; \alpha^n) = [g(z_t; \alpha^n) - g(z_t; \alpha^0)] + g(z_t; \alpha^0)$. We obtain:

$$[T^{-1} \textstyle\sum_{t=1}^{T} g^2(z_t; \alpha^n)]^{\frac{1}{2}} \leqq [T^{-1} Q_T(\alpha^n)]^{\frac{1}{2}} + [T^{-1} \sum_{t=1}^{T} g^2(z_t; \alpha^0)]^{\frac{1}{2}}, \tag{49}$$

which contradicts Assumption 8 because $\{\alpha^n\}$ is unbounded whereas the right-hand term of (49) is, for arbitrarily large T, bounded by a number G independent of T. Indeed, the first term is bounded by $(4\sigma^2 + \partial)^{\frac{1}{2}}$ as a consequence of (48), whereas the second one is also bounded because of the compactness of Z and the continuity of $g(z; \alpha^0)$.

In order to prove the second part of Assumption 4, let us now assume K is any compact set. By Assumption 6 the function $[g(z;\alpha) - g(z; \beta)]^2$ is continuous on the compact set $Z \times K \times K$. Hence,

$$\varphi_T(\alpha, \beta) = \textstyle\int_Z [g(z; \alpha) - g(z; \beta)]^2 \, d\mu_T(z) \tag{50}$$

converges to:

$$\varphi(\alpha, \beta) = \textstyle\int_Z [g(z; \alpha) - g(z; \beta)]^2 \, d\mu(z). \tag{51}$$

Assumption 7 implies that $\varphi(\alpha, \beta)$ vanishes only when $\alpha = \beta$. To complete the proof of Assumption 4 we need only show that on $K \times K$ the function $\varphi(\alpha, \beta)$ is continuous and the convergence of $\varphi_T(\alpha, \beta)$ to $\varphi(\alpha, \beta)$ uniform. But continuity of $g(z; \alpha)$ implies continuity of $\varphi_T(\alpha, \beta)$, from which continuity of $\varphi(\alpha, \beta)$ will follow when uniform convergence has been proven.

The uniform convergence of φ_T to φ results from Theorem 1 of P. Billingsley and F. Topsøe (1967), of which the following proposition is a direct corollary.[7]

Let $\mathcal{F}$ *be a family of real functions* f *on* Z. *Let* $\mathcal{F}$ *be equicontinuous and bounded. Then* $\int f d\mu_T$ *converges uniformly to* $\int f d\mu$ *for every sequence of measures* μ_T *on* Z *that converges weakly to* μ.

To prove uniform convergence of φ_T to φ we apply this proposition, with f equal to $[g(z; \alpha) - g(z; \beta)]^2$, considered as a function of z, and with the family $\mathcal{F}$ made of the set of all f such that (α, β) belongs to $K \times K$. The continuous function g is bounded on the compact set $Z \times K$; hence the family $\mathcal{F}$ is bounded. Equicontinuity

[7] This proposition and the reference to Billingsley and Topsøe were given to me by W. Hildenbrand.

means that for every $\varepsilon > 0$ there exists δ such that $||z-w|| < \delta$ implies $|f(z)-f(w)| < \varepsilon$ for all $f \in \mathscr{F}$. Equicontinuity of $\mathscr{F}$ is implied here by the continuity of g and the compactness of $Z \times K \times K$. (Indeed, assume equicontinuity does not hold. There is then some $\varepsilon > 0$ and, for $n = 1, 2 \cdots$ *ad infinitum*, sequences δ^n, z^n, w^n, α^n, β^n such that δ^n decreases to zero, $||z^n - w^n|| < \delta^n$ and $|f^n(z^n) - f^n(w^n)| > \varepsilon$, f^n corresponding to (α^n, β^n). There are, however, subsequences that simultaneously converge to 0, $z^0 = w^0$, α^0 and β^0. But continuity of g implies $|f^0(z^0) - f^0(w^0)| < \varepsilon$, a contradiction.) This completes the proof of Theorem 3.

6. Multivariate regressions. Let us consider briefly the multivariate generalization of model (1), namely:

$$x_{it} = g_i(z_{1t} \cdots z_{mt}; \alpha_1 \cdots \alpha_p) + \varepsilon_{it} \qquad i = 1,2 \cdots n. \tag{52}$$

Regression estimates will now be obtained as minimizing a quadratic form of the deviations from the x_i to the corresponding g_i.

More precisely let x_t, $g_t(\alpha)$ and ε_t be the n-vectors with components x_{it}, $g_i(z_t; \alpha)$ and ε_{it}. Let S_T be any, possibly random, positive definite square matrix. Finally, let $\hat{\alpha}(S_T)$ be the vector minimizing:[8]

$$L_T(\alpha, S_T) = \sum_{t=1}^{T} [x_t - g_t(\alpha)]' S_T [x_t - g_t(\alpha)]. \tag{53}$$

We are interested in the conditions under which $\hat{\alpha}(S_T)$ tends to the true value α^0.

One may check that, if S_T tends in probability to a positive definite matrix S, the consistency proofs listed above essentially carry through with minor changes. For instance, the notation introduced at the beginning of Section 2 is naturally changed as follows. From the vectors $q_t(\alpha)$ defined by (9) the quadratic form $Q_T(\alpha)$ is computed as:

$$Q_T(\alpha) = \sum_{t=1}^{T} q_t(\alpha)' S_T q_t(\alpha). \tag{54}$$

The vectors $\lambda_{tT}(\alpha)$ are then defined by (11) and the numbers $u_T(\alpha)$ by:

$$u_T(\alpha) = \sum_{t=1}^{T} \lambda_{tT}(\alpha)' S_T \varepsilon_t. \tag{55}$$

The lemma of Section 2 then stands after condition (i) is replaced by:

(i) $\Pr\{\inf_{\alpha \in \omega} Q_T(\alpha) = 0\}$ tends to zero as T increases indefinitely.

In Assumption 1 the variance σ^2 must be replaced by the covariance matrix Σ, and in the constrained linear model the vector $a(\alpha)$ replaced by a matrix $A(\alpha)$. Except for these obvious changes, Theorem 1 still holds.

Similar changes are required in order to transpose Theorem 2 and Theorem 3.

---

[8] In practice S_T will often be an estimate of the inverse of the covariance matrix Σ of ε_t. For instance, in view of defining an asymptotically efficient procedure, one may consider the following computations: (i) derive $\hat{\alpha}(I)$ taking S_T as being the identity matrix, (ii) compute $\hat{\varepsilon}_t = x_t - g_t[\hat{\alpha}(I)$ and $\hat{M}_{\varepsilon\varepsilon} = T^{-1} \sum_{t=1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t'$, (iii) estimate α by $\hat{\alpha}(\hat{M}_{\varepsilon\varepsilon}^{-1})$.

Assumptions 4 to 8 read as before except that $g(z; \alpha)$ is now the n-vector with components $g_i(z; \alpha)$ and that (28), (29), and (47) are replaced respectively by:

$$\phi_T(\alpha, \beta) = T^{-1} \sum_{t=1}^{T} [g_t(\alpha) - g_t(\beta)]' S[g_t(\alpha) - g_t(\beta)] \tag{28'}$$

$$\phi_T(\alpha, \alpha^0) \geqq 4 \operatorname{tr} S\Sigma + \partial \tag{29'}$$

$$T^{-1} \sum_{t=1}^{T} g(z_t; \alpha)' S g(z_t; \alpha) \leqq G. \tag{47'}$$

Acknowledgments. This article was discussed by many friends, in particular by R. Radner and T. Rothenberg. I benefited from suggestions made after my presentation of the paper at the Cowles Foundation and at Harvard University. The constructive comments of a referee and the consultation given by W. Hildenbrand about uniform weak convergence of measures were very helpful.

REFERENCES

[1] BILLINGSLEY, P. and TOPSØE, F. (1967). Uniformity in weak convergence. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **7** 1–16.

[2] FISHER, I. (1937). Note on a short-cut method for calculating distributed lags. *Bull. Internat. Statist.* **29**.

[3] JENNRICH, R. I. (1969). Asymptotic properties of non-linear least-squares estimators. *Ann. Math. Statist.* **40** 633–643.

[4] MALINVAUD, E. (1966). *Statistical Methods of Econometrics*. North-Holland Publishing Company, translation of a book first published in French in 1964 by Dunod.