

STATISTICAL INFERENCE BASED ON M-ESTIMATORS FOR THE
MULTIVARIATE NONLINEAR REGRESSION MODEL IN IMPLICIT FORM

by Geraldo Souza and A. Ronald Gallant

North Carolina State University

SUMMARY

The general regression model postulates that an observed multivariate response is a function of an observed multivariate input, an unknown parameter, and an unobservable additive multivariate error. In principle one may solve for the response given the input, parameter, and error but this is not required in applications. Given an optimization procedure which defines an estimator, a companion theory of large sample inference is developed. This theory includes strong consistency and asymptotic normality of the estimator and the asymptotic null and non-null distributions of the Wald test statistic, an analog of the likelihood ratio test statistic, and an analog of Rao's efficient score test statistic.

AMS 1970 subject classifications. Primary 62F05, 62H15; Secondary 62G35, 62P20.

Key words and phrases. Multivariate nonlinear regression, M-estimators, nonlinear simultaneous equation models, Wald test, Rao's efficient score test, likelihood ratio test, non-null asymptotic distributions.

Abbreviated Title. Tests from M-estimators in multivariate nonlinear regression.

1. Introduction. An M-variate response y_t follows the statistical model

$$q(y_t, x_t, \gamma^\circ) = e_t \quad t = 1, 2, \dots, n$$

where x_t denotes a k-dimensional input variable, γ° denotes an unknown s-dimensional parameter, and e_t denotes an M-variate random error. These variables are contained in the Borel sets \mathcal{U} , \mathcal{X} , Γ , and \mathcal{E} respectively. It is convenient to absorb the scale parameters of the error distribution into γ and impose

ASSUMPTION 1. The errors e_t are independently and identically distributed each with mean zero and variance-covariance matrix the identity.

An example, which occurs in the study of consumer demand (Jorgenson, Christenson and Lau, 1975), is

$$y_{1t} = \frac{\theta_1 + \theta_2 \ln(p_1/I)_t + \theta_3 \ln(p_2/I)_t + (\theta_6 - \theta_2 - \theta_3) \ln(p_3/I)_t}{-1 + \theta_6 \ln(p_1/I)_t + \theta_7 \ln(p_2/I)_t + \theta_8 \ln(p_3/I)_t} = e_{1t}$$

$$y_{2t} = \frac{\theta_4 + \theta_3 \ln(p_1/I)_t + \theta_5 \ln(p_2/I)_t + (\theta_7 - \theta_2 - \theta_5) \ln(p_3/I)_t}{-1 + \theta_6 \ln(p_1/I)_t + \theta_7 \ln(p_2/I)_t + \theta_8 \ln(p_3/I)_t} = e_{2t}$$

or, say,

$$y_{1t} - f_1(x_t, \theta) = e_{1t}$$

$$y_{2t} - f_2(x_t, \theta) = e_{2t}$$

with $x_t = (\ln p_1, \ln p_2, \ln p_3, \ln I)_t$.^{1/} In this model, termed a translog expenditure system in the econometric literature, y_{1t} and y_{2t} are the t^{th} consumer's expenditures on non-durable goods and services, of durable goods expressed as a proportion of the consumer's income I_t ; p_1 , p_2 , and p_3 are the prices of non-durable goods, services of durable goods, and services respectively. The scale parameters may be absorbed by writing

$$q(x_t, \gamma) = \begin{pmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{pmatrix} \begin{pmatrix} y_{1t} - f_1(x_t, \theta) \\ y_{2t} - f_2(x_t, \theta) \end{pmatrix}$$

or, say,

$$q(x_t, \gamma) = R[y - f(x_t, \theta)]$$

where $\gamma = (\theta_1, \dots, \theta_8, r_{11}, r_{12}, r_{22})$ with $r_{11}, r_{22} > 0$.

The models envisaged here are supposed to describe the behavior of a physical, biological, economic, or social system. If so, to each value of (e, x, γ) there should correspond one and only one outcome y . This condition and continuity are imposed.

ASSUMPTION 2. For each $(x, \gamma) \in \mathcal{X} \times \Gamma$ the equation $q(y, x, \gamma) = e$ defines a one-to-one mapping of \mathcal{E} into \mathcal{U} denoted as $Y(e, x, \gamma)$. Moreover, $Y(e, x, \gamma)$ is continuous on $\mathcal{E} \times \mathcal{X} \times \Gamma$.

Throughout, $q(y, x, \gamma) = e$ will be referred to as the structural model while $y = Y(e, x, \gamma)$ will be termed the reduced form following the conventions of the econometric literature. It should be emphasized that it is not necessary to have a closed form expression for the reduced form, or even to be able to compute it using numerical methods, in order to use the statistical methods set forth here.

Interest is focused on a p -vector of parameters λ . Typically, λ will equal γ or some subvector of γ . The parameter λ is contained in Λ and is estimated by finding that value $\hat{\lambda}_n$ in Λ which maximizes

$$s_n(\lambda) = (1/n) \sum_{t=1}^n s(y_t, x_t, \hat{\tau}_n, \lambda)$$

where $\hat{\tau}_n$ is a random variable; typically, τ corresponds to some subvector of γ which is regarded as a nuisance parameter and $\hat{\tau}_n$ is its estimator.

This formulation of the estimation problem is motivated by a consideration of the statistical methods presently in use in nonlinear regression analysis and some others one might wish to employ. The translog example may be used for illustration.

Maximum likelihood methods for nonlinear models with explicit, separable reduced forms - those which may be written as $y = f(x, \theta) + e$ - have

been studied by Barnett (1976) and Holly (1978). The translog example fits this description and if normally distributed errors are assumed then the log likelihood is

$$\text{const} + n \ln \det R - (1/2) \sum_{t=1}^n \|R[y_t - f(x_t, \theta)]\|^2 .$$

Put $\lambda = (\theta_1, \dots, \theta_8, r_{11}, r_{12}, r_{22})$ and

$$s(y, x, \tau, \lambda) = \ln \det R - (1/2) \|R[y - f(x, \theta)]\|^2$$

and the method of maximum likelihood may be formulated as above; the dependence of $s(y, x, \tau, \lambda)$ on τ is trivial in this case.

The estimation method which probably finds most frequent use in applications is "iterated Aitken" also termed "Zellner-type," "the seemingly unrelated regression method," and "minimum distance." This method has been studied in Gallant (1975) and Holly (1978). The method is as follows when applied to the translog. First least squares residuals \hat{u}_{1t} and \hat{u}_{2t} are obtained by fitting the two models $y_{1t} = f_1(x_t, \theta) + u_{1t}$ and $y_{2t} = f_2(x_t, \theta) + u_{2t}$ individually by least squares. Let

$$\hat{S} = (1/n) \sum_{t=1}^n \begin{pmatrix} \hat{u}_{1t} \\ \hat{u}_{2t} \end{pmatrix} (\hat{u}_{1t}, \hat{u}_{2t}) .$$

The iterated Aitken estimator is obtained by finding $\hat{\theta}$ which minimizes

$$(1/n) \sum_{t=1}^n [y_t - f(x_t, \theta)]' \hat{S}^{-1} [y_t - f(x_t, \theta)] .$$

Set $\lambda = (\theta_1, \dots, \theta_8)$, $\tau = (s_{11}, s_{12}, s_{22})$, and

$$s(y, x, \tau, \lambda) = -(1/2) [y_t - f(x_t, \theta)]' S^{-1} [y_t - f(x_t, \theta)]$$

and the iterated Aitken estimator is seen to be of the general form considered here.

The asymptotic properties of the estimator $\hat{\lambda}_n$ are considered in Sections 3 and 4. To indicate the nature of these results, for the moment assume that a density $p(y)$ and Jacobian $(\partial/\partial y')q(y, x, \gamma)$ are available so that a conditional density for the endogeneous variables y is given by,

$$p(y|x, \gamma) = |\det(\partial/\partial y')q(y, x, \gamma)| p[q(y, x, \gamma)].$$

Set the true value $\gamma^\circ = \gamma^*$ and suppose that $\lim_{n \rightarrow \infty} \hat{\tau}_n = \tau^*$ almost surely. Then $\hat{\lambda}_n$ converges almost surely to the point λ^* which maximizes

$$\bar{s}(\gamma^*, \tau^*, \lambda) = \lim_{n \rightarrow \infty} (1/n) \sum_{t=1}^n \int_U s(y, x_t, \tau^*, \gamma) p(y|x_t, \gamma^*) dy.$$

Moreover, $\sqrt{n}(\hat{\lambda}_n - \lambda^*)$ converges in distribution to a multivariate normal $N(0, V)$. A strongly consistent estimator of V is $\hat{V} = \hat{J}^{-1} \hat{J} \hat{J}^{-1}$ where \hat{J}

$$\hat{J} = (1/n) \sum_{t=1}^n [(\partial/\partial \lambda) s(y_t, x_t, \hat{\tau}_n, \hat{\lambda}_n)] [(\partial/\partial \lambda) s(y_t, x_t, \hat{\tau}_n, \hat{\lambda}_n)]'$$

$$\hat{J} = -(1/n) \sum_{t=1}^n (\partial^2/\partial \lambda \partial \lambda') s(y_t, x_t, \hat{\tau}_n, \hat{\lambda}_n).$$

This result suggest that if the true value is γ_n° in each finite sample then the estimator $\hat{\lambda}_n$ is to be regarded as estimating that point λ_n° which maximizes $\bar{s}(\gamma_n^\circ, \tau^*, \lambda)$. Typically, λ and τ are subvectors of γ whence there correspond λ_n° and τ_n° obtained as subvectors of γ_n° . Usually, τ_n° is regarded as a nuisance parameter to be held fixed at $\tau_n^\circ = \tau^*$ for all n . This implies a constraint on γ_n° . It may happen that λ_n° obtained as a subvector of γ_n° does not coincide with λ_n° defined as the maximum of $\bar{s}(\gamma_n^\circ, \tau^*, \lambda)$. In these cases, the definition of λ_n° as a maximum is to prevail. Similarly, the definition of τ^* as the almost sure limit of $\hat{\tau}_n$ is to prevail when ambiguities arise. The results of Section 3 and 4 are obtained with γ_n° subject to drift to facilitate the derivation of non-null distributions in Section 5.

Inference is considered as an adjunct to the estimation procedure in Section 5. To indicate the nature of these results, consider testing

$$H: h(\lambda_n^0) = 0 \text{ against } A: h(\lambda_n^0) \neq 0$$

where $h(\lambda)$ is an r -vector valued function with Jacobian $H(\lambda) = (\partial/\partial\lambda')h(\lambda)$; let $\hat{h} = h(\hat{\lambda}_n)$ and $\hat{H} = H(\hat{\lambda}_n)$. Following Wald, the test statistic

$$W = n \hat{h}'(\hat{H} \nabla \hat{H}')^{-1} \hat{h}$$

may be employed to test H against A . Let $\tilde{\lambda}_n$ denote an estimator obtained by maximizing $s_n(\lambda)$ subject to the constraint $h(\lambda) = 0$. By analogy with the likelihood ratio test, the test statistic

$$L = -2n[s_n(\tilde{\lambda}_n) - s_n(\hat{\lambda}_n)]$$

may be employed to test H against A. Let $\tilde{H} = H(\tilde{\lambda}_n)$.

$$\tilde{J} = -(1/n) \sum_{t=1}^n (\partial^2 / \partial \lambda \partial \lambda') s(y_t, x_t, \hat{\tau}_n, \tilde{\lambda}_n),$$

$$\tilde{J} = (1/n) \sum_{t=1}^n [(\partial / \partial \lambda) s(y_t, x_t, \hat{\tau}_n, \tilde{\lambda}_n) \{ (\partial / \partial \lambda) s(y_t, x_t, \hat{\tau}_n, \tilde{\lambda}_n) \}'],$$

and $\tilde{V} = \tilde{J}^{-1} \tilde{J} \tilde{J}^{-1}$. The analogy with Rao's efficient score test statistic is

$$R = n [(\partial / \partial \lambda) s_n(\tilde{\lambda}_n)]' \tilde{J}^{-1} \tilde{H}' (\tilde{H} \tilde{V}^{-1} \tilde{H}')^{-1} \tilde{H} \tilde{J}^{-1} [(\partial / \partial \lambda) s_n(\tilde{\lambda}_n)].$$

These test statistics are shown to converge in distribution to the same non-central chi-square distribution with r degrees of freedom. The non-centrality parameter of this non-central chi-square distribution may be approximated by

$$\alpha_n^{\circ} = n h'(\lambda_n^{\circ}) [H(\lambda_n^{\circ}) V^{\circ} H'(\lambda_n^{\circ})]^{-1} h(\lambda_n^{\circ}) / 2$$

where $V^{\circ} = (J^{\circ})^{-1} J^{\circ} (J^{\circ})^{-1}$ and

$$J^{\circ} = (1/n) \sum_{t=1}^n \int_U \{ (\partial / \partial \lambda) s(y, x_t, \tau_n^{\circ}, \lambda_n^{\circ}) \} \{ (\partial / \partial \lambda) s(y, x_t, \tau_n^{\circ}, \lambda_n^{\circ}) \}' p(y | x_t, \gamma_n^{\circ}) dy$$

$$J^{\circ} = -(1/n) \sum_{t=1}^n \int_U (\partial^2 / \partial \lambda \partial \lambda') s(y, x_t, \tau_n^{\circ}, \lambda_n^{\circ}) p(y | x_t, \gamma_n^{\circ}) dy.$$

The validity of this approximation requires the additional assumption that

$$\lim_{n \rightarrow \infty} (J^{\circ} - J^{\circ}) = 0 \text{ in the case of the statistic } L.$$

2. The Probability Space and Limits of Cesaro Sums. Repeatedly, in the sequel, the almost sure uniform limit of a Cesaro sum such as

$$(1/n) \sum_{t=1}^n f(y_t, x_t, \gamma) = (1/n) \sum_{t=1}^n f[Y(e_t, x_t, \gamma^0), x_t, \gamma]$$

is required. In the nonlinear regression literature much attention has been devoted to finding conditions which insure this behavior yet are plausible and can be easily recognized as obtaining or not obtaining in an application (Jennrich, 1969; Malinvaud, 1970; Gallant, 1977; Gallant and Holly, 1978).

A précis of these ideas appears here for the readers convenience.

The independent variables (x_1, x_2, \dots) are either fixed, the realization of a random process or the components of the vectors x_t are a mixture of these. If, say, the data for the translog example were generated by randomly selecting individuals from a population in each of several years then a plausible and convenient assumption is that $x_t = (\ln p_1, \ln p_2, \ln p_3, \ln I)_t$ follows some distribution which is absolutely continuous with respect to Lebesgue measure on the cube X . Then, denoting this distribution by $\mu(x)$,

$$\begin{aligned} \lim_{n \rightarrow \infty} (1/n) \sum_{t=1}^n f(p_{1t}, p_{2t}, p_{3t}, I_t) \\ = \int f(p_1, p_2, p_3, I) d\mu(x) \end{aligned}$$

for integrable f (strong law of large numbers). The typical regression assumption is that the independent variables x_t and the errors e_t are uncorrelated. If this is strengthened to independence then, by Assumption 1,

$$\begin{aligned} \lim_{n \rightarrow \infty} (1/n) \sum_{t=1}^n f(p_{1t}, p_{2t}, p_{3t}, I_t, e_t) \\ = \iint f(p_1, p_2, p_3, I, e) dP(e) d\mu(x) \end{aligned}$$

for integrable f where P is the distribution of the errors.

These considerations motivate the notion of regarding the joint sequence of independent variables and errors as being defined on a probability space

and the definition of Cesaro summability. Theorem 1 demonstrates that the concept is not void for examples which come immediately to mind, and Theorem 2 yields the desired uniform, almost sure convergence.

DEFINITION. (Gallant and Holly, 1978) A sequence $\{v_t\}$ of points from a Borel set V is said to generate Cesaro summable sequences with respect to a probability measure ν defined on the Borel subsets of V and a dominating function $b(v)$ with $\int b d\nu < \infty$ if

$$\lim_{n \rightarrow \infty} (1/n) \sum_{t=1}^n f(v_t) = \int f(v) d\nu(v)$$

for every real valued, continuous function f with $|f(v)| < b(v)$.

THEOREM 1. (Gallant and Holly, 1978) Let V_t , $t = 1, 2, \dots$ be a sequence of independent and identically distributed s -dimensional random variables defined on a complete probability space $(\Omega, \mathcal{a}, P^*)$ with common distribution ν . Let ν be absolutely continuous with respect to some product measure on R^s and let b be a non-negative function with $\int b d\nu < \infty$. Then there exists E with $P^*(E) = 0$ such that if $\omega \notin E$

$$\lim_{n \rightarrow \infty} (1/n) \sum_{t=1}^n f[V_t(\omega)] = \int f(v) d\nu(v)$$

for every continuous function with $|f(v)| \leq b(v)$.

Note that the null set depends on b and not on f .

THEOREM 2. (Gallant, 1977) Let $f(v, \lambda)$ be a real valued continuous function on $V \times \Lambda$ where Λ is compact. Let $\{v_t\}$ generate Cesaro summable sequences with respect to ν and $b(v)$. If $\sup_{\Lambda} |f(v, \lambda)| < b(v)$ then

$$\lim_{n \rightarrow \infty} \sup_{\Lambda} \left| (1/n) \sum_{t=1}^n f(v_t, \lambda) - \int f(v, \lambda) d\nu(v) \right| = 0$$

and $\int f(v, \lambda) d\nu$ is continuous on Λ .

ASSUMPTION 3. (Gallant and Holly, 1978) Almost every realization of $\{v_t\}$ with $v_t = (e_t, x_t)$ generates Cesaro summable sequences with respect to the product measure

$$\nu(A) = \int_{\mathcal{X}} \int_{\mathcal{E}} I_A(e, x) dP(e) d\mu(x)$$

and a dominating function $b(e, x)$. Almost every sequence $\{x_t\}$ generates Cesaro summable sequences with respect to μ and $b(x) = \int_{\mathcal{E}} b(e, x) dP(e)$. For each $x \in \mathcal{X}$ there is a neighborhood N_x such that $\int_{\mathcal{E}} \sup_{N_x} b(e, x) dP(e) < \infty$.

COROLLARY. (Gallant and Holly, 1978) Let Assumptions 1 through 3 hold. Let $f(y, x, \rho)$ be continuous on $U \times \mathcal{X} \times K$ where K is compact. Let

$|f(y, x, \rho)| \leq b[q(y, x, \gamma^0), x]$ for all $(y, x) \in U \times \mathcal{X}$ and all (ρ, γ^0) in $K \times \Lambda^*$ where Λ^* is compact. Then both $(1/n) \sum_{t=1}^n f(y_t, x_t, \rho)$ and $(1/n) \sum_{t=1}^n \int_{\mathcal{E}} f[Y(e, x_t, \gamma^0), x_t, \rho] dP(e)$ converge uniformly to

$$\int_{\mathcal{X}} \int_{\mathcal{E}} f[Y(e, x, \gamma^0), x, \rho] dP(e) d\mu(x)$$

except on the event E with $P^*(E) = 0$ given by Assumption 3.

In typical applications, a density $p(e)$ and a Jacobian

$$J(y, x, \gamma^0) = (\partial/\partial y') q(y, x, \gamma^0)$$

are available. With these in hand, the conditional density

$$p(y|x, \gamma^0) = |\det J(y, x, \gamma^0)| p[q(y, x, \gamma^0)]$$

may be used for computing limits since

$$\begin{aligned} & \int_{\mathcal{X}} \int_{\mathcal{E}} f[Y(e, x, \gamma^0), x, \gamma] dP(e) d\mu(x) \\ &= \int_{\mathcal{X}} \int_U f(y, x, \gamma) p(y|x, \gamma^0) dy d\mu(x) \end{aligned}$$

The choice of integration formulas is dictated by convenience.

The probabilistic structure which is usually assumed in asymptotic regression analysis is as follows. One fixes a sequence

$$x_{\infty} = (x_1, x_2, \dots)$$

and works with a sequence of regular conditional probability distributions $P_n(\cdot | x_\infty, \gamma^\circ)$ defined on the measurable subsets of $\mathcal{Y}_n = \prod_{t=1}^n \mathcal{Y}$. The assumption of independent and identically distributed errors with common distribution $P(\cdot)$, Assumption 1, induces a product measure $P_\infty(\cdot)$ defined on the measurable subsets of $\mathcal{E}_\infty = \prod_{t=1}^\infty \mathcal{E}$. The relationship between the conditional distribution on \mathcal{Y}_n and the error distribution on \mathcal{E}_∞ is

$$P_n(A | x_\infty, \gamma^\circ) = P_\infty\{e_\infty \in \mathcal{E}_\infty : [Y(e_1, x_1, \gamma^\circ), \dots, Y(e_n, x_n, \gamma^\circ)] \in A\}$$

for every measurable subset of \mathcal{Y}_n . A statement such as $\hat{\lambda}_n$ converges almost surely to λ^* means that $\hat{\lambda}_n$ is a random variable with argument $\frac{3}{(e_1, \dots, e_n, x_1, \dots, x_n, \gamma^\circ)}$ and that $P_\infty(E) = 0$ where

$$E = \bigcap_{\epsilon > 0} \bigcap_{j=1}^\infty \bigcup_{n=j}^\infty \{e_\infty : |\hat{\lambda}_n - \lambda^*| > \epsilon\} .$$

A statement that $\sqrt{n}(\hat{\lambda}_n - \lambda^*)$ converges in distribution to a multivariate normal distribution $N(\cdot | \delta, V)$ means that for A of the form

$$A = (-\infty, \lambda_1] \times (-\infty, \lambda_2] \times \dots \times (-\infty, \lambda_r]$$

it is true that

$$\lim_{n \rightarrow \infty} P_n(A | x_\infty, \gamma^\circ) = \int_A dN(z | \delta, V) .$$

In the sequel, the usual conventions will be followed; the probabilistic structure is as described in the preceding paragraph. The analysis is conditional on a fixed sequence x_∞ for which the Cesaro summability property holds with respect to μ and $b(x)$ of Assumption 3. The link with the usual probabilistic structure and Assumption 3 is provided by Theorem 3 below. It shows that the event $E \subset \mathcal{E}_\infty$ on which Cesaro summability fails for the joint sequence (e_1, x_1) , (e_2, x_2) , ... occurs with P_∞ probability zero for

almost every choice of x_∞ .

THEOREM 3. Let Assumption 1 hold and let $\{V_t\}$ with $V_t = (E_t, X_t)$ be a sequence of random variables defined on $(\Omega, \mathcal{G}, P^*)$ which satisfies Assumption 3. Let P_∞ be the measure on the measurable subsets of $\mathcal{E}_\infty = \prod_{t=1}^\infty \mathcal{E}$ induced by the sequence of random variables $\{E_t\}$ and, similarly, μ_∞ on \mathcal{X}_∞ induced by $\{X_t\}$. Then there is a subset N of \mathcal{X}_∞ with $\mu_\infty(N) = 0$ such that each $x_\infty^\circ \notin N$ both satisfies the Cesaro summability property (with respect to $b(x)$ and μ) and

$$E^\circ = \bigcap_{\epsilon > 0} \bigcap_{j=0}^\infty \bigcup_{n=j}^\infty \{e_\infty : \mathbb{E} |f(e, x)| < b(e, x) \ni |(1/n) \sum_{t=1}^n f(e_t, x_t^\circ) - \iint f dP d\mu| > \epsilon\}$$

has $P_\infty(E^\circ) = 0$.

PROOF. By Assumption 3, $\mu_\infty(N_1) = 0$ where

$$N_1 = \bigcap_{\epsilon > 0} \bigcap_{j=0}^\infty \bigcup_{n=j}^\infty \{x_\infty : \mathbb{E} |f(x)| < b(x) \ni |(1/n) \sum_{t=1}^n f(x_t) - \int f d\mu| > \epsilon\}.$$

Moreover, Assumption 3 implies that $P_\infty \times \mu_\infty(F) = 0$ where

$$F = \bigcap_{\epsilon > 0} \bigcap_{j=0}^\infty \bigcup_{n=j}^\infty \{(e_\infty, x_\infty^\circ) : \mathbb{E} |f(e, x)| < b(e, x) \ni |(1/n) \sum_{t=1}^n f(e_t, x_t) - \iint f dP d\mu| > \epsilon\}$$

Note that for every x_∞° , $E^\circ \times \{x_\infty^\circ\} \subset F$ whence

$$\begin{aligned} P_\infty(E^\circ) &= \int_{\mathcal{E}_\infty} I_{E^\circ}(e_\infty) dP_\infty(e_\infty) \\ &= \int_{\mathcal{E}_\infty} I_{E^\circ \times \{x_\infty^\circ\}}(e_\infty, x_\infty^\circ) dP_\infty(e_\infty) \\ &\leq \int_{\mathcal{E}_\infty} I_F(e_\infty, x_\infty^\circ) dP_\infty(e_\infty). \end{aligned}$$

But

$$P_\infty \times \mu_\infty(F) = \int_{\mathcal{X}_\infty} \int_{\mathcal{E}_\infty} I_F(e_\infty, x_\infty^\circ) dP_\infty(e_\infty) d\mu_\infty(x_\infty^\circ)$$

whence $P_\infty(E^\circ) = 0$ except for x_∞° in some event N_2 with $\mu_\infty(N_2) = 0$. Let

$$N = N_1 \cup N_2. \quad \square$$

3. Consistency. It is necessary to introduce a dependence of the true parameter on sample size in order to derive the non-null asymptotic distribution of test statistics. Also, the large sample behavior of $\hat{\tau}_n$ must be specified.

NOTATION.

$$s_n(\lambda) = (1/n) \sum_{t=1}^n s(y_t, x_t, \hat{\tau}_n, \lambda)$$

$$\bar{s}(\gamma, \tau, \lambda) = \int_{\mathcal{X}} \int_{\mathcal{E}} s[Y(e, x, \gamma), x, \tau, \lambda] dP(e) d\mu(x)$$

ASSUMPTION 4. The parameter γ° is indexed by n and $\lim_{n \rightarrow \infty} \gamma_n^\circ = \gamma^*$ for some point $\gamma^* \in \Gamma$. The sequence $\hat{\tau}_n$ converges almost surely to a point τ^* and $\sqrt{n}(\hat{\tau}_n - \tau^*)$ is bounded in probability.^{4/} There are unique points $\lambda^*, \lambda_1^\circ, \lambda_2^\circ, \dots$ corresponding to $\gamma = \gamma^*, \gamma_1^\circ, \gamma_2^\circ, \dots$ which maximize $\bar{s}(\gamma, \tau^*, \lambda)$ over Λ . The function $h(\lambda)$ of the hypothesis $H: h(\lambda_n^\circ) = 0$ is a continuous vector valued function on Λ ; the point λ^* satisfies $h(\lambda^*) = 0$ and $\lim_{n \rightarrow \infty} \sqrt{n}(\lambda_n^\circ - \lambda^*) = \delta$.

To illustrate, consider the translog example with the iterated Aitken estimation method. In this case,

$$\bar{s}(\gamma_n^\circ, \tau^*, \lambda) = -M/2 - \frac{1}{2} \int_{\mathcal{X}} [f(x, \theta) - f(x, \theta_n^\circ)]' \Sigma^{-1} [f(x, \theta) - f(x, \theta_n^\circ)] d\mu(x) .$$

where $\Sigma = (R^{-1})(R^{-1})'$ provided τ^* is such that $S^* = S \Big|_{\tau=\tau^*} = \Sigma$. It is seen

at sight that

$$\lambda_n^\circ = (\theta_1, \theta_2, \dots, \theta_8)_n^\circ$$

maximizes $\bar{s}(\gamma_n^\circ, \tau^*, \lambda)$. Uniqueness obtains if Σ is positive definite and $\mu(A) > 0$ when $\theta \neq \theta^\circ$ where

$$A = \{x: f(x, \theta) \neq f(x, \theta^\circ)\} > 0 .$$

Recall that $x = (\ln p_1, \ln p_2, \ln p_3, \ln I) \in \mathcal{X}$ and μ is absolutely continuous with respect to Lebesgue measure on \mathcal{X} whence $\mu(A) = 0$ if and only if $f(x, \theta) = f(x, \theta^\circ)$ a.e. with respect to Lebesgue measure on \mathcal{X} . Some algebraic manipulation shows that $f(x, \theta) = f(x, \theta^\circ)$ a.e. implies $\theta = \theta^\circ$ provided that, for example, $\theta_1^\circ, \theta_4^\circ$ are non-zero and at least one of the values $\theta_6^\circ, \theta_7^\circ$, or θ_8° is non-zero.^{5/} Thus, λ_n° is uniquely determined for the translog example with minimum distance estimation.

The almost sure convergence imposed in Assumption 4 implies that there is a sequence which takes its values in a neighborhood of τ^* and is tail equivalent to $\hat{\tau}_n$. Thus, without loss of generality, it may be assumed that $\hat{\tau}_n$ takes its values in a compact sphere T for which τ^* is an interior point. Similarly, Γ may be taken as a compact sphere with interior point γ^* . Sufficient conditions such that Λ may effectively be taken as a compact sphere are set forth in Theorem 4; they are patterned after Huber (1964).

THEOREM 4. Let Assumptions 1 through 4 hold. Suppose that Γ and T are compact, that Λ is an unbounded closed set, and that there is a continuous positive function $w(\lambda)$ such that

- (i) $\sup_{\Lambda} s[Y(e, x, \gamma), x, \tau, \lambda] / w(\lambda)$ is continuous and $\sup_{\Lambda} |s[Y(e, x, \gamma), x, \tau, \lambda] / w(\gamma)|$ is dominated by $b(e, x)$
- (ii) $\int_{\mathcal{X}} \int_{\mathcal{E}} \sup_{\Lambda} \{s[Y(e, x, \gamma^*), x, \tau^*, \lambda] / w(\lambda)\} \leq -1$
- (iii) $\liminf_{\|\lambda\| \rightarrow \infty} w(\lambda) > -\bar{s}(\gamma^*, \tau^*, \lambda^*)$.

Then there is a compact set Λ' containing γ^* such that there corresponds to almost every realization of $\{x_t\}$ an N for which $n > N$ implies

$$\sup_{\Lambda'} s_n(\lambda) = \sup_{\Lambda} s_n(\lambda);$$

N depends on the realization but Λ' does not.

The function $s(y, x, \tau, \lambda)$ is continuous on $U \times X \times T \times \Lambda'$ and $s(y, x, \tau, \lambda) \leq b[q(y, x, \gamma), x]$ on $U \times X \times T \times \Lambda' \times \Gamma$. (The function $b(e, x)$ is given by Assumption 3.)

THEOREM 5. (Strong consistency) Let Assumptions 1 through 5 hold. Then $\hat{\lambda}_n$ and $\tilde{\lambda}_n$ converge almost surely to λ^* .

PROOF. Let $\{e_t\}$ be such that $\lim_{n \rightarrow \infty} \hat{\tau}_n = \tau^*$, $\{(e_t, x_t)\}$ has the Cesaro summability property, and $\hat{\lambda}_n \in \Lambda'$ for large n ; almost every error sequence is such. Since λ^* satisfies $h(\lambda^*) = 0$ and $\lambda^* \in \Lambda'$ it follows that $\tilde{\lambda}_n \in \Lambda'$ for large n . Since Λ' is compact, the sequences $\hat{\lambda}_n$ and $\tilde{\lambda}_n$ corresponding to $\{e_t\}$ have subsequences $\hat{\lambda}_{n_m}$ and $\tilde{\lambda}_{n_m}$ converging to limit points $\dot{\lambda}$ and $\ddot{\lambda}$ respectively. By the Corollary of Theorem 2,

$$s_n(\gamma, \tau, \lambda) = (1/n) \sum_{t=1}^n s[Y(e_t, x_t, \gamma), x_t, \tau, \lambda]$$

converges uniformly to $\bar{s}(\gamma, \tau, \lambda)$ on $\Gamma \times T \times \Lambda'$. Then

$$\begin{aligned} \bar{s}(\gamma^*, \tau^*, \dot{\lambda}) &= \lim_{n \rightarrow \infty} s_{n_m}(\gamma_{n_m}^{\circ}, \hat{\tau}_{n_m}, \hat{\lambda}_{n_m}) \\ &\geq \lim_{m \rightarrow \infty} s_{n_m}(\gamma_{n_m}^{\circ}, \hat{\tau}_{n_m}, \lambda^*) \\ &= \bar{s}(\gamma^*, \tau^*, \lambda^*) \end{aligned}$$

because in each finite sample $(\gamma_{n_m}^{\circ}, \hat{\tau}_{n_m}, \hat{\lambda}_{n_m})$ maximizes $s_{n_m}(\gamma_{n_m}^{\circ}, \hat{\tau}_{n_m}, \lambda)$ while $(\gamma_{n_m}^{\circ}, \hat{\tau}_{n_m}, \lambda^*)$ need not. Similarly, because λ^* satisfies $h(\lambda^*) = 0$, $\bar{s}(\gamma^*, \tau^*, \ddot{\lambda}) \geq \bar{s}(\gamma^*, \tau^*, \lambda^*)$. The assumption of a unique maximum, Assumption 4, implies $\dot{\lambda} = \ddot{\lambda} = \lambda^*$. Then $\{\hat{\lambda}_n\}$ and $\{\tilde{\lambda}_n\}$ have only the one limit point λ^* . \square

4. Asymptotic Normality. The asymptotic normality of $\hat{\lambda}_n$ is established here. The verification that the "scores" $(\partial/\partial\lambda)s(y_t, x_t, \hat{\tau}_n, \lambda_n^{\circ})$ are asymptotically normally distributed is the critical result; the notion of Cesaro summable sequences plays a key role in the proof. From this result

PROOF. By (iii) there is a compact set Λ' and an ϵ with $0 < \epsilon < 1$ such that

$$\inf_{\lambda \notin \Lambda'} w(\lambda) > \frac{-\bar{s} + \epsilon}{1 - \epsilon}$$

where $\bar{s} = \bar{s}(\gamma^*, \tau^*, \lambda^*) < 0$. By (i), (ii) and the Corollary of Theorem 2, there exists N_1 such that for $n > N$.

$$\sup_{\lambda \notin \Lambda'} s_n(\lambda)/w(\lambda) \leq \sup_{\lambda \in \Lambda} s_n(\lambda)/w(\lambda) \leq -1 + \epsilon$$

Then $\lambda \notin \Lambda'$ and $n > N_1$ imply

$$s_n(\lambda) \leq -w(\lambda)(1 - \epsilon) \leq \bar{s} - \epsilon.$$

By (i) and the Corollary of Theorem 2,

$$\lim_{n \rightarrow \infty} s_n(\lambda^*)/w(\lambda^*) = \bar{s}/w(\lambda^*)$$

so there is an N_2 such that $n > N_2$ implies

$$s_n(\lambda^*) > \bar{s} - \epsilon.$$

Then if $n > N = \max\{N_1, N_2\}$ it follows that $\lambda^* \in \Lambda'$ and

$$\sup_{\lambda \in \Lambda} s_n(\lambda) > \bar{s} - \epsilon \geq \sup_{\lambda \notin \Lambda'} s_n(\lambda). \quad \square$$

Theorem 4 suggests that an assumption that it eventually suffices to maximize $s_n(\lambda)$ over a compact set is not as restrictive as might appear at first glance. As an alternative to Theorem 4, one may verify this assumption directly as in, say, Gallant and Holly (1978).

ASSUMPTION 5. The sets Γ and T are compact spheres containing γ^* and τ^* , respectively. There is a compact set Λ' containing λ^* and to almost every realization of $\{e_t\}$ there corresponds an N for which $n > N$ implies

$$\sup_{\Lambda'} s_n(\lambda) = \sup_{\Lambda} s_n(\lambda).$$

asymptotic normality of $\hat{\lambda}_n$ follows by Taylor's series expansions and related arguments of the sort which are typical of nonlinear regression theory.

ASSUMPTION 6. There are open spheres Γ^* , T^* , and Λ^* with $\gamma^* \in \Gamma^* \subset \Gamma$, $\tau^* \in T^* \subset T$, and $\lambda^* \in \Lambda^* \subset \Lambda'$. The elements of $(\partial/\partial\lambda)s(y, s, \tau, \lambda)$, $(\partial^2/\partial\lambda\partial\lambda')s(y, x, \tau, \lambda)$, $(\partial^2/\partial\tau\partial\lambda')s(y, x, \tau, \lambda)$, and $[(\partial/\partial\lambda)s(y, x, \tau, \lambda)] \times [(\partial/\partial\lambda)s(y, x, \tau, \lambda)]'$ are continuous and dominated by $b[q(y, x, \gamma), x]$ on $\mathcal{Y} \times \mathcal{X} \times \bar{\Gamma}^* \times \bar{T}^* \times \bar{\Lambda}^*$ where the overbar indicates the closure of a set. Moreover,

$$\int_{\mathcal{E}} (\partial/\partial\lambda)s[Y(e, x, \gamma_n^0), x, \tau^*, \lambda_n^*] dP(e) = 0$$

$$\int_{\mathcal{X}} \int_{\mathcal{E}} (\partial^2/\partial\tau\partial\lambda')s[Y(e, x, \gamma^*), x, \tau^*, \lambda^*] dP(e) d\mu(x) = 0 .$$

The two integral conditions imposed in Assumption 6 do not appear to impede application of the results in instances which come readily to mind. Apparently they are intrinsic properties of reasonable estimation procedures. The translog example with iterated Aitken estimation serves to illustrate. The score is

$$(\partial/\partial\lambda)s(y, x, \tau, \lambda) = [(\partial/\partial\theta')f(x, \theta)]' s^{-1} [y - f(x, \theta)]$$

whence

$$(\partial/\partial\lambda)s[Y(e, x, \gamma_n^0), x, \tau^*, \lambda_n^*] = [(\partial/\partial\theta')f(x, \theta)]' (s^*)^{-1} e .$$

Since $\int_{\mathcal{E}} e dP(e) = 0$, one notes at sight that the two integral conditions are satisfied.

NOTATION.

$$\mathcal{J} = \int_{\mathcal{X}} \int_{\mathcal{E}} \{(\partial/\partial\lambda)s[Y(e, x, \gamma^*), x, \tau^*, \lambda^*]\} \{(\partial/\partial\lambda)s[Y(e, x, \gamma^*), x, \tau^*, \lambda^*]\}' dP(e) d\mu(x)$$

$$\mathcal{g} = - \int_{\mathcal{X}} \int_{\mathcal{E}} (\partial^2/\partial\lambda\partial\lambda')s[Y(e, x, \gamma^*), x, \tau^*, \lambda^*] dP(e) d\mu(x)$$

$$\mathcal{J}_n(\lambda) = (1/n) \sum_{t=1}^n [(\partial/\partial\lambda)s(y_t, x_t, \hat{\tau}_n, \lambda)] [(\partial/\partial\lambda)s(y_t, x_t, \hat{\tau}_n, \lambda)]'$$

$$\mathcal{J}_n(\lambda) = -(1/n) \sum_{t=1}^n (\partial^2/\partial\lambda\partial\lambda') s(y_t, x_t, \hat{\tau}_n, \lambda) \quad .$$

For the translog example with iterated Aitken estimation

$$\mathcal{J} = \int_{\mathcal{X}} [(\partial/\partial\theta')f(x, \theta)]' (S^*)^{-1} \Sigma (S^*)^{-1} [(\partial/\partial\theta')f(x, \theta)] d\mu(x)$$

$$\mathcal{J} = \int_{\mathcal{X}} [(\partial/\partial\theta')f(x, \theta)]' (S^*)^{-1} [(\partial/\partial\theta')f(x, \theta)] d\mu(x) \quad .$$

If $\Sigma = S^*$ then $\mathcal{J} = \mathcal{J}$.

THEOREM 6. (Asymptotic Normality of the Scores) Under Assumptions 1 through 6

$$(1/\sqrt{n}) \sum_{t=1}^n (\partial/\partial\lambda)s(y_t, x_t, \hat{\tau}_n, \lambda_n^0) \xrightarrow{\mathcal{L}} N(0, \mathcal{J}) \quad ,$$

\mathcal{J} may be singular.

PROOF. Given l with $\|l\| = 1$ consider the triangular array of random variables

$$Z_{tn} = l' (\partial/\partial\lambda)s[Y(e_t, x_t, \gamma_n^0), x_t, \tau^*, \lambda_n^0] \quad t = 1, \dots, n; n = 1, 2, \dots \quad .$$

Each Z_{tn} has mean, $\int_{\mathcal{E}} Z_{tn}(e) dP(e)$, zero by assumption and variance

$$\sigma_{tn}^2 = l' \int_{\mathcal{E}} \{(\partial/\partial\lambda)s[Y(e, x_t, \gamma_n^0), x_t, \tau^*, \lambda_n^0]\} \{(\partial/\partial\lambda)s[Y(e, x_t, \gamma_n^0), x_t, \tau^*, \lambda_n^0]\}' dP(e) l \quad .$$

By the Corollary of Theorem 2 and the assumption that $\lim_{n \rightarrow \infty} (\gamma_n^0, \lambda_n^0) = (\gamma^*, \lambda^*)$

it follows that $\lim_{n \rightarrow \infty} (1/n)V_n = l' \mathcal{J} l$ where

$$V_n = \sum_{t=1}^n \sigma_{tn}^2 \quad .$$

Now $(1/n)V_n$ is the variance of $(1/\sqrt{n}) \sum_{t=1}^n Z_{tn}$ and if $l' \mathcal{J} l = 0$ then $(1/\sqrt{n}) \sum_{t=1}^n Z_{tn}$ converges in distribution to $N(0, l' \mathcal{J} l)$ by Chebyshev's inequality.

Suppose, then, that $l' \mathcal{J} l > 0$. If it is shown that for every $\epsilon > 0$

$\lim_{n \rightarrow \infty} B_n = 0$ where

$$B_n = (1/n) \sum_{t=1}^n \int \mathbb{I}_{\mathcal{E}} [|z| > \epsilon \sqrt{V_n}] [Z_{tn}(e)] Z_{tn}^2(e) dP(e)$$

then $\lim_{n \rightarrow \infty} (n/V_n) B_n = 0$. This is the Lindberg-Feller condition (Chung, 1974);

it implies that $(1/\sqrt{n}) \sum_{t=1}^n Z_{tn}$ converges in distribution to $N(0, \ell' \mathcal{J} \ell)$.

Let $n > 0$ and $\epsilon > 0$ be given. Choose $a > 0$ such that $\bar{B}(\gamma^*, \lambda^*) < n/2$

where

$$\begin{aligned} \bar{B}(\gamma^*, \lambda^*) = & \int \int \mathbb{I}_{\mathcal{E}} [|z| > \epsilon a] \{ \ell'(\partial/\partial \lambda) s[Y(e, x, \gamma^*), x, \tau^*, \lambda^*] \} \\ & \times \{ \ell'(\partial/\partial \lambda) s[Y(e, x, \gamma^*), x, \tau^*, \lambda^*] \}^2 dP(e) d\mu(x) . \end{aligned}$$

This is possible because $\bar{B}(\gamma^*, \lambda^*)$ exists when $a = 0$. Choose a continuous function $\Theta(z)$ and an N_1 such that, for all $n > N_1$,

$$\mathbb{I}_{\mathcal{E}} [|z| > \epsilon \sqrt{V_n}](z) \leq \Theta(z) \leq \mathbb{I}_{\mathcal{E}} [|z| > \epsilon a](z)$$

and set

$$\begin{aligned} \tilde{B}_n(\gamma, \lambda) = & (1/n) \sum_{t=1}^n \int \mathcal{E} \Theta \{ \ell'(\partial/\partial \lambda) s[Y(e, x, \gamma), x_t, \tau^*, \lambda] \} \\ & \times \{ \ell'(\partial/\partial \lambda) s[Y(e, x, \gamma), x_t, \tau^*, \lambda] \}^2 dP(e) . \end{aligned}$$

By the Corollary of Theorem 2, $\tilde{B}_n(\gamma, \lambda)$ converges uniformly on $\bar{\Gamma}^* \times \bar{\Lambda}^*$ to, say, $\tilde{B}(\gamma, \lambda)$. By assumption $\lim_{n \rightarrow \infty} (\gamma_n^\circ, \lambda_n^\circ) = (\gamma^*, \lambda^*)$ whence $\lim_{n \rightarrow \infty} \tilde{B}_n(\gamma_n^\circ, \lambda_n^\circ) = \tilde{B}(\gamma^*, \lambda^*)$. Then there is an N_2 such that, for all $n > N_2$, $\tilde{B}_n(\gamma_n^\circ, \lambda_n^\circ) < \tilde{B}(\gamma^*, \lambda^*) + n/2$.

But, for all $n > N = \max\{N_1, N_2\}$, $B_n \leq \tilde{B}_n(\gamma_n^\circ, \lambda_n^\circ)$ whence

$$B_n \leq \tilde{B}_n(\gamma_n^\circ, \lambda_n^\circ) < \tilde{B}(\gamma^*, \lambda^*) + n/2 \leq \bar{B}(\gamma^*, \lambda^*) + n/2 < n .$$

By Taylor's theorem

$$\begin{aligned} (1/\sqrt{n}) \sum_{t=1}^n Z_{tn} = & (1/\sqrt{n}) \ell' \sum_{t=1}^n (\partial/\partial \lambda) s(y_t, x_t, \hat{\tau}_n^\circ, \lambda_n^\circ) \\ & + [(1/n) (\partial/\partial \tau') \ell' \sum_{t=1}^n (\partial/\partial \lambda) s(y_t, x_t, \bar{\tau}_n, \lambda_n^\circ)] \sqrt{n} (\hat{\tau}_n - \tau^*) \end{aligned}$$

where $\|\bar{\tau}_n - \tau^*\| \leq \|\hat{\tau}_n - \tau^*\|$. ^{5/} By the Corollary of Theorem 2, the almost sure convergence of $\hat{\tau}_n$, and Assumption 5, the vector multiplying $\sqrt{n}(\hat{\tau}_n - \tau^*)$ converges almost surely to zero. This and the assumed probability bound on $\sqrt{n}(\hat{\tau}_n - \tau^*)$ imply that the last term converges in probability to zero whence $(1/\sqrt{n})\ell' \sum_{t=1}^n (\partial/\partial \lambda) s(y_t, x_t, \hat{\tau}_n, \lambda_n^0) \xrightarrow{\mathcal{L}} N(0, \ell' \mathcal{J} \ell)$. This holds for every ℓ with $\|\ell\| = 1$ whence the desired result obtains. \square

THEOREM 7. Let Assumptions 1 through 6 hold. Then \mathcal{J} is nonsingular,

$$(1/\sqrt{n}) \sum_{t=1}^n (\partial/\partial \lambda) s(y_t, x_t, \hat{\tau}_n, \lambda^*) \xrightarrow{\mathcal{L}} N(\mathcal{J}\delta, \mathcal{J}),$$

$$\sqrt{n}(\hat{\lambda}_n - \lambda^*) \xrightarrow{\mathcal{L}} N(\delta, \mathcal{J}^{-1} \mathcal{J} \mathcal{J}^{-1}),$$

$\mathcal{J}_n(\hat{\lambda}_n)$ converges almost surely to \mathcal{J} , and $\mathcal{J}_n(\hat{\lambda}_n)$ converges almost surely to \mathcal{J} .

PROOF. By the mean value theorem and the dominated convergence theorem, interchange of differentiation and integration is permitted whence

$$\mathcal{J} = -(\partial^2/\partial \lambda \partial \lambda') \bar{s}(y^*, \tau^*, \lambda^*).$$

By Assumption 4, $\bar{s}(y^*, \tau^*, \lambda)$ has a unique maximum at $\lambda = \lambda^*$ whence $(\partial^2/\partial \lambda \partial \lambda') \bar{s}(y^*, \tau^*, \lambda)$ must be negative definite at $\lambda = \lambda^*$.

By Taylor's theorem,

$$\begin{aligned} & (1/\sqrt{n}) \sum_{t=1}^n (\partial/\partial \lambda) s(y_t, x_t, \hat{\tau}_n, \lambda^*) \\ &= (1/\sqrt{n}) \sum_{t=1}^n (\partial/\partial \lambda) s(y_t, x_t, \hat{\tau}_n, \lambda_n^0) - \bar{\mathcal{J}} \sqrt{n}(\lambda^* - \lambda_n^0) \end{aligned}$$

where $\bar{\mathcal{J}}$ has rows

$$- (1/n) \sum_{t=1}^n (\partial/\partial \lambda') (\partial/\partial \lambda_i) s[Y(e_t, x_t, \gamma_n^0), x_t, \hat{\tau}_n, \bar{\lambda}_{in}]$$

and $\|\bar{\lambda}_{in} - \lambda^*\| \leq \|\lambda^* - \lambda_n^0\|$. From the Corollary of Theorem 2 and the assumption

that $(\gamma_n^\circ, \hat{\tau}_n, \lambda_n^\circ)$ converges almost surely to $(\gamma^*, \tau^*, \lambda^*)$ it follows that \bar{g} converges almost surely to g . Since $\lim_{n \rightarrow \infty} \sqrt{n}(\lambda_n^\circ - \lambda^*) = \delta$ by Assumption, the left hand side converges in distribution to $N(g, \delta, \mathcal{J})$.

By Theorem 5 there is a sequence which is tail equivalent to $\hat{\lambda}_n$ and takes its values in Λ^* . The remarks apply to the tail equivalent sequence but a new notation is not introduced. Taylor's theorem applies to this sequence whence

$$\begin{aligned} & (1/\sqrt{n}) \sum_{t=1}^n (\partial/\partial \lambda) s(y_t, x_t, \hat{\tau}_n, \lambda^*) \\ &= (1/\sqrt{n}) \sum_{t=1}^n (\partial/\partial \lambda) s(y_t, x_t, \hat{\tau}_n, \hat{\lambda}_n) - \bar{g} \sqrt{n} (\lambda^* - \hat{\lambda}_n) \end{aligned}$$

where \bar{g} is similar to \bar{g} of the previous paragraph and converges almost surely to g for similar reasons. Now the first term on the right is the gradient of the objective function evaluated at a random variable which is tail equivalent to the optimizing random variable; it must, therefore, converge almost surely to zero. Thus, $\sqrt{n}(\hat{\lambda}_n - \lambda^*) \xrightarrow{\mathcal{L}} N(\delta, g^{-1} \mathcal{J} g^{-1})$ by Slutsky's theorem.

By the Corollary of Theorem 2 and the almost sure convergence of $(\gamma_n^\circ, \hat{\tau}_n, \hat{\lambda}_n)$ to $(\gamma^*, \tau^*, \lambda^*)$ it follows that $\lim_{n \rightarrow \infty} [\mathcal{J}_n(\hat{\lambda}_n), \mathcal{J}_n(\hat{\lambda}_n)] = (\mathcal{J}, \mathcal{J})$ almost surely. \square

5. Tests of Hypotheses. Tests of the hypothesis

$$H: h(\lambda^\circ) = 0 \quad \text{against} \quad A: h(\lambda^\circ) \neq 0$$

are considered here. A full rank assumption is imposed which is not strictly necessary. However, the less than full rank case appears to be of no practical importance and a full rank assumption eliminates much clutter from the theorems and proofs.

NOTATION

$$\tilde{\lambda}_n \text{ maximizes } s_n(\lambda) \text{ subject to } h(\lambda) = 0$$

$$\mathfrak{J} = \mathcal{J}_n(\hat{\lambda}_n), \quad \tilde{\mathfrak{J}} = \mathcal{J}_n(\tilde{\lambda}_n)$$

$$\hat{g} = g_n(\hat{\lambda}_n), \quad \tilde{g} = g_n(\tilde{\lambda}_n)$$

$$V = g^{-1} g g^{-1}, \quad \hat{V} = \hat{g}^{-1} \hat{g} \hat{g}^{-1}, \quad \tilde{V} = \tilde{g}^{-1} \tilde{g} \tilde{g}^{-1}$$

$$H(\lambda) = (\partial/\partial\lambda')h(\lambda) \quad (\text{the Jacobian of } h \text{ of order } r \times p)$$

$$h = h(\lambda^*), \quad \hat{h} = h(\hat{\lambda}_n), \quad \tilde{h} = h(\tilde{\lambda}_n), \quad h^0 = h(\lambda_n^0)$$

$$H = H(\lambda^*), \quad \hat{H} = H(\hat{\lambda}_n), \quad \tilde{H} = H(\tilde{\lambda}_n), \quad H^0 = H(\lambda_n^0)$$

ASSUMPTION 7. The r -vector valued function $h(\lambda)$ defining the hypothesis $H: h(\lambda^0) = 0$ is twice continuously differentiable with Jacobian $H(\lambda) = (\partial/\partial\lambda')h(\lambda)$; $H(\lambda)$ has full rank at $\lambda = \lambda^*$. The matrix $V = g^{-1} g g^{-1}$ has full rank. The statement "the null hypothesis is true" means that $h(\lambda_n^0) = 0$ for all n .

THEOREM 8. Under Assumptions 1 through 7 the statistics

$$W = n h'(\hat{\lambda}_n)(\hat{H}\hat{V}\hat{H}')^{-1}h(\hat{\lambda}_n)$$

$$R = n[(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)]\tilde{g}^{-1}\tilde{H}'(\tilde{H}\tilde{V}\tilde{H}')^{-1}\tilde{H}\tilde{g}^{-1}[(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)]$$

converge in distribution to the non-central Chi square distribution with r degrees of freedom and noncentrality parameter α

$$\alpha = \delta'H'(H V H')^{-1}H\delta/2$$

Under the null hypothesis, the limiting distribution is the central chi square with r degrees of freedom.

PROOF. (The statistic W) By Theorem 5 there is a sequence which is tail equivalent to $\hat{\lambda}_n$ and takes its values in Λ^* . The remarks refer to the tail equivalent sequence but a new notation is not introduced. Taylor's theorem applies to this sequence whence

$$\sqrt{n}[h_i(\hat{\lambda}_n) - h_i(\lambda^*)] = (\partial/\partial\lambda')h_i(\bar{\lambda}_{in})\sqrt{n}(\hat{\lambda}_n - \lambda^*) \quad i = 1, 2, \dots, r$$

where $\|\bar{\lambda}_{in} - \lambda^*\| \leq \|\hat{\lambda}_n - \lambda^*\|$. By Theorem 5, $\lim_{n \rightarrow \infty} \|\bar{\lambda}_{in} - \lambda^*\| = 0$ almost surely whence $\lim_{n \rightarrow \infty} (\partial/\partial\lambda)h_i(\bar{\lambda}_{in}) = (\partial/\partial\lambda)h_i(\lambda^*)$ almost surely. Now, in addition, $h(\theta^*) = 0$ so the Taylor's expansion may be written^{8/}

$\sqrt{n} h(\hat{\lambda}_n) = [H + o_s(1)]\sqrt{n}(\hat{\lambda}_n - \lambda^*)$. Then by Theorem 7, $\sqrt{n} h(\hat{\lambda}_n)$ has the same asymptotic distribution as $H\sqrt{n}(\hat{\lambda}_n - \lambda^*)$. Now $(\hat{H} \hat{V} \hat{H}')^{-\frac{1}{2}}$ exists for n sufficiently large and converges almost surely to $(H V H')^{-\frac{1}{2}}$ whence $(\hat{H} \hat{V} \hat{H}')^{-\frac{1}{2}}\sqrt{n} h(\hat{\lambda}_n)$ and $(H V H')^{-\frac{1}{2}}H\sqrt{n}(\hat{\lambda}_n - \lambda^*)$ have the same asymptotic distribution. But

$$(H V H')^{-\frac{1}{2}}H\sqrt{n}(\hat{\lambda}_n - \lambda^*) \xrightarrow{d} N[(H V H')^{-\frac{1}{2}}H\delta, I_r]$$

whence W converges in distribution to the non-central chi-square.

When the null hypothesis is true, it follows from Taylor's theorem that

$$0 = \sqrt{n}[h_i(\lambda_n^0) - h_i(\lambda^*)] = [(\partial/\partial\lambda')h_i(\bar{\lambda}_{in})]\sqrt{n}(\lambda_n^0 - \lambda^*)$$

Taking the limit as n tends to infinity this equation becomes

$$0 = (\partial/\partial\lambda')h_i(\lambda^*)\delta \quad \text{whence } H\delta = 0 \text{ and } \lambda = 0$$

(The statistic R) By Theorem 5 there is a sequence which is tail equivalent to $\tilde{\lambda}_n$ and takes its values in Λ^* . The remarks below refer to the tail equivalent sequence but a new notation is not introduced. By Taylor's theorem

$$(\partial/\partial\lambda_i)s_n(\tilde{\lambda}_n) = (\partial/\partial\lambda_i)s_n(\lambda^*) + [(\partial^2/\partial\lambda\partial\lambda_i)s_n(\bar{\lambda}_{in})]'(\tilde{\lambda}_n - \lambda^*)$$

$$h_i(\tilde{\lambda}_n) = h_i(\lambda^*) + [(\partial/\partial\lambda')h_i(\bar{\lambda}_{in})](\tilde{\lambda}_n - \lambda^*)$$

where $\|\bar{\lambda}_{in} - \lambda^*\|$, $\|\bar{\lambda}_{in} - \lambda^*\| \leq \|\tilde{\lambda}_n - \lambda^*\|$ for $i = 1, 2, \dots, r$. By tail equivalence, there is for every realization of $\{e_t\}$ an N such that $h(\tilde{\lambda}_n) = 0$ for all $n > N$. Thus $h(\tilde{\lambda}_n) = o_s(1)$ and recall that $h(\theta^*) = 0$. Then the continuity of $H(\lambda)$, the almost sure convergence of $\tilde{\lambda}_n$ to λ^* given by Theorem 5, and the Corollary of Theorem 2 permit these Taylor's expansions to be rewritten as

$$(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) = (\partial/\partial\lambda)s_n(\lambda^*) - [\mathcal{J} + o_s(1)](\tilde{\lambda}_n - \lambda^*)$$

$$[H + o_s(1)](\tilde{\lambda}_n - \lambda^*) = o_s(1) \quad .$$

These equations may be reduced algebraically to

$$[H + o_s(1)][\mathcal{J} + o_s(1)]^{-1}(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) = [H + o_s(1)][\mathcal{J} + o_s(1)]^{-1}(\partial/\partial\lambda)s_n(\lambda^*) + o_s(1) \quad .$$

Then it follows from Theorem 7 that

$$[H + o_s(1)][\mathcal{J} + o_s(1)]^{-1}\sqrt{n}(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) \xrightarrow{\mathcal{L}} N(H\delta, HVH') \quad .$$

The continuity of $H(\lambda)$, Theorem 5, and the Corollary of Theorem 2 permit the conclusion that

$$(\tilde{H} \tilde{V} \tilde{H}')^{-\frac{1}{2}} \tilde{H} \tilde{\mathcal{J}}^{-1} \sqrt{n}(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) \xrightarrow{\mathcal{L}} N[(HVH')^{-\frac{1}{2}}H\delta, I_r]$$

whence R converges in distribution to the non-central chi square. This completes the argument but note for the next proof that

$$H \mathcal{J}^{-1} \sqrt{n} (\partial/\partial\lambda)s_n(\tilde{\lambda}_n) \xrightarrow{\mathcal{L}} N(H\delta, HV^{-1}H') \quad . \quad \square$$

THEOREM 9. Under Assumptions 1 through 6 the statistic

$$L = -2[s_n(\tilde{\lambda}_n) - s_n(\hat{\lambda}_n)]$$

converges in distribution to the law of the quadratic form

$$Y = Z' \mathcal{J} Z$$

where Z is distributed as the multivariate normal

$$Z \sim N[\mathcal{J}^{-1} H' (H \mathcal{J}^{-1} H')^{-1} H \delta, \mathcal{J}^{-1} H' (H \mathcal{J}^{-1} H')^{-1} (H V H') (H \mathcal{J}^{-1} H')^{-1} H \mathcal{J}^{-1}] .$$

If $\mathcal{J} = \mathcal{J}$ then Y has the non-central chi-square distribution with r degrees of freedom and non-centrality parameter

$$\alpha = \delta' H' (H V H')^{-1} H \delta / 2 .$$

Under the null hypothesis Y is distributed as the central chi-square with r degrees of freedom provided that $\mathcal{J} = \mathcal{J}$.

PROOF. By Theorem 5 there are sequences which are tail equivalent to $\hat{\lambda}_n$ and $\tilde{\lambda}_n$ and take their values in Λ^* . The remarks below refer to the tail equivalent sequences but a new notation is not introduced. By Taylor's theorem

$$\begin{aligned} & -2n[s_n(\tilde{\lambda}_n) - s_n(\hat{\lambda}_n)] \\ & = -2n[(\partial/\partial\lambda)s_n(\hat{\lambda}_n)]'(\tilde{\lambda}_n - \hat{\lambda}_n) - n(\tilde{\lambda}_n - \hat{\lambda}_n)'[(\partial^2/\partial\lambda\partial\lambda')s_n(\tilde{\lambda}_n)](\tilde{\lambda}_n - \hat{\lambda}_n) \end{aligned}$$

where $\|\tilde{\lambda}_n - \hat{\lambda}_n\| \leq \|\tilde{\lambda}_n - \hat{\lambda}_n\|$. The Corollary of Theorem 2 and the almost sure convergence of $(\tilde{\lambda}_n, \hat{\lambda}_n)$ to (λ^*, λ^*) imply that $(\partial^2/\partial\lambda\partial\lambda')s_n(\tilde{\lambda}_n) = -[\mathcal{J} + o_s(1)]$.

Now, by tail equivalence, $-2n(\partial/\partial\lambda')s_n(\hat{\lambda}_n) = o_s(1)$ whence

$$-2n[s_n(\tilde{\lambda}_n) - s_n(\hat{\lambda}_n)] = n(\tilde{\lambda}_n - \hat{\lambda}_n)'[\mathcal{J} + o_s(1)](\tilde{\lambda}_n - \hat{\lambda}_n) + o_s(1) .$$

By tail equivalence, there is for every realization of $\{e_t\}$ an N such that $\tilde{\lambda}_n$ solves

$$\text{Max } s_n(\lambda) \quad \text{subject to } h(\lambda) = 0$$

for all $n > N$. Then for $n > N$ there are Lagrange multipliers θ_n such that

$$\sqrt{n}(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) - H'(\tilde{\lambda}_n) \sqrt{n} \theta_n = 0 .$$

Thus,

$$[H + o_s(1)]' \sqrt{n} \theta_n = \sqrt{n} (\partial/\partial \lambda) s_n(\tilde{\lambda}_n) + o_s(1)$$

by a previous argument $\sqrt{n}(\partial/\partial \lambda) s_n(\hat{\lambda}_n) = o_s(1)$ whence

$$= \sqrt{n} (\partial/\partial \lambda) s_n(\tilde{\lambda}_n) - \sqrt{n} (\partial/\partial \lambda) s_n(\hat{\lambda}_n) + o_s(1)$$

by Taylor's theorem and previous arguments

$$= [g + o_s(1)] \sqrt{n} (\hat{\lambda}_n - \tilde{\lambda}_n) + o_s(1) .$$

From this string of equalities one has

$$H g^{-1} [H + o_s(1)]' \sqrt{n} \theta_n = \sqrt{n} H g^{-1} (\partial/\partial \lambda) s_n(\tilde{\lambda}_n) + o_s(1)$$

whence by the last line of the previous proof

$$H g^{-1} [H + o_s(1)]' \sqrt{n} \theta_n \xrightarrow{\mathcal{L}} N(H\delta, H V^{-1} H') .$$

Thus

$$\sqrt{n} \theta \xrightarrow{\mathcal{L}} N[(H g^{-1} H')^{-1} H \delta, (H g^{-1} H')^{-1} (H V^{-1} H') (H g^{-1} H')^{-1}] .$$

Again from the string of equalities one has

$$g^{-1} [H + o_s(1)]' \sqrt{n} \theta_n = g^{-1} [g + o_s(1)] \sqrt{n} (\hat{\lambda}_n - \tilde{\lambda}_n) + o_s(1)$$

whence

$$\sqrt{n} (\hat{\lambda}_n - \tilde{\lambda}_n) \xrightarrow{\mathcal{L}} N[g^{-1} H' (H g^{-1} H')^{-1} H \delta, g^{-1} H (H g^{-1} H')^{-1} (H V^{-1} H') (H g^{-1} H')^{-1} H' g^{-1}] .$$

Then $\sqrt{n}(\hat{\lambda}_n - \tilde{\lambda}_n)$ converges in distribution to the distribution of the random variable Z and $\sqrt{n}(\hat{\lambda}_n - \tilde{\lambda}_n) = O_P(1)$. From the first paragraph of the proof,

$$\begin{aligned} & -2n[s_n(\tilde{\lambda}_n) - s_n(\hat{\lambda}_n)] \\ &= n(\hat{\lambda}_n - \tilde{\lambda}_n)' [g + o_s(1)] (\hat{\lambda}_n - \tilde{\lambda}_n) + o_s(1) \\ &= n(\hat{\lambda}_n - \tilde{\lambda}_n)' g (\hat{\lambda}_n - \tilde{\lambda}_n) + O_P(1) o_s(1) O_P(1) + o_s(1) . \end{aligned}$$

If $\mathcal{J} = \mathcal{J}$ then $V = \mathcal{J}^{-1}$ and

$$Z \sim N[\mathcal{J}^{-1}H'(H\mathcal{J}^{-1}H')^{-1}H\delta, \mathcal{J}^{-1}H'(H\mathcal{J}^{-1}H')^{-1}H\mathcal{J}^{-1}] .$$

The conclusion that Y is chi-square follows at once from Theorem 2 of Searle (1971, p. 57). \square

In a typical application, λ and τ are subvectors of γ or some easily computed function of γ . Thus, if γ_n° is specified then λ_n° and τ_n° become specified and this λ_n° will satisfy Assumption 4. Thus, in a typical application, $(\gamma_n^\circ, \tau_n^\circ, \lambda_n^\circ)$ is specified and the noncentrality parameter $\alpha = \delta'H'(HVH')^{-1}H\delta/2$ is to be computed. The annoyance of having to specify $(\gamma^*, \tau^*, \lambda^*)$ in order to make this computation may be eliminated by application of Theorem 10.

NOTATION.

$$\mathcal{J}^\circ = (1/n) \sum_{t=1}^n \int_{\mathcal{E}} \{(\partial/\partial\lambda)s[Y(e, x_t, \gamma_n^\circ), x_t, \tau_n^\circ, \lambda_n^\circ]\} \\ \times \{(\partial/\partial\lambda)s[Y(e, x_t, \gamma_n^\circ), x_t, \tau_n^\circ, \lambda_n^\circ]\}' dP(e)$$

$$\mathcal{J}^\circ = -(1/n) \sum_{t=1}^n \int_{\mathcal{E}} (\partial^2/\partial\lambda\partial\lambda')s[Y(e, x_t, \gamma_n^\circ), x_t, \tau_n^\circ, \lambda_n^\circ] dP(e)$$

$$V^\circ = (\mathcal{J}^\circ)^{-1} \mathcal{J}^\circ (\mathcal{J}^\circ)^{-1}$$

$$\alpha_n^\circ = n h'(\lambda_n^\circ)[H(\lambda_n^\circ) V^\circ H'(\lambda_n^\circ)]^{-1} h(\lambda_n^\circ)/2$$

THEOREM 10. Let Assumptions 1 through 7 hold and let $\{\gamma_n^\circ, \tau_n^\circ, \lambda_n^\circ\}$ be any sequence with $\lim_{n \rightarrow \infty} (\gamma_n^\circ, \tau_n^\circ, \lambda_n^\circ) = (\gamma^*, \tau^*, \lambda^*)$ and $\lim_{n \rightarrow \infty} \sqrt{n}(\lambda_n^\circ - \lambda^*) = \delta$. Then $\lim_{n \rightarrow \infty} \alpha_n^\circ = \alpha$.

PROOF. By the continuity of $H(\lambda)$, the Corollary of Theorem 2, and the assumption that $\lim_{n \rightarrow \infty} (\gamma_n^\circ, \tau_n^\circ, \lambda_n^\circ) = (\gamma^*, \tau^*, \lambda^*)$ it follows that

$\lim_{n \rightarrow \infty} [H(\lambda_n^\circ) V^\circ H'(\lambda_n^\circ)]^{-1} = (HVH')^{-1}$. By Taylor's theorem

$$\begin{aligned} \sqrt{n} h_i(\lambda_n^\circ) &= \sqrt{n} [h_i(\lambda_n^\circ) - h_i(\lambda^*)] \\ &= [(\partial/\partial \lambda) h_i(\bar{\lambda}_{in})]' \sqrt{n} (\lambda_n^\circ - \lambda^*) \end{aligned}$$

where $\|\bar{\lambda}_{in} - \lambda^*\| \leq \|\lambda_n^\circ - \lambda^*\|$ for $i = 1, 2, \dots, r$. Thus,

$$\lim_{n \rightarrow \infty} \sqrt{n} h(\lambda_n^\circ) = H\delta \quad \square$$

FOOTNOTES

1/ See Caves and Christensen (1978) for a detailed discussion of the domain of applicability of this demand system. Since the purpose here is illustrative, a simple expedient is adopted: let Θ and \mathcal{X} be compact cubes in R^8 and R^4 such that on $\Theta \times \mathcal{X}$

$$-1 + \theta_6 \ln(p_1/I) + \theta_7 \ln(p_2/I) + \theta_8 \ln(p_3/I) > 0 \quad .$$

2/ For a real valued function $g(u,v)$ of the k -vector u and p -vector v , $(\partial/\partial u)g(u,v)$ denotes the k -vector $[(\partial/\partial u_1)g(u,v), \dots, (\partial/\partial u_k)g(u,v)]'$, and $(\partial^2/\partial u \partial v')$ $g(u,v)$ denotes the $k \times p$ matrix with typical element $(\partial^2/\partial u_i \partial v_j)g(u,v)$. For a k -vector valued function $g(v)$ of the p -vector v , $(\partial/\partial v')g(v)$ denotes the $k \times p$ matrix with typical element $(\partial/\partial v_j)g_j(v)$. For a $k \times p$ vector matrix function $g(z)$, $\int g(u) d\mu(u)$ denotes the $k \times p$ matrix with typical element $\int g_{ij}(u) d\mu(u)$.

3/ Typically $\hat{\lambda}_n$ depends on $(y_1, \dots, y_n, x_1, \dots, x_n)$ and the dependence on $(e_1, \dots, e_n, \gamma_n^\circ)$ enters via the relation $y_t = Y(e_t, x_t, \gamma_n^\circ)$.

4/ Given $\delta > 0$ there exists M and N such that $P(\sqrt{n} \|\hat{\tau}_n - \tau^*\| < M) > 1 - \delta$ for all $n > N$.

- 5/ The weakest condition is that a particular matrix has full rank but a display of this matrix would consume too much space.
- 6/ The function $\bar{\tau}_n$ may be taken as measurable; however, measurability of $\bar{\tau}_n$ is not necessary for the validity of the proof.
- 7/ Searle's (1971) definition of the non-central chi-square is used.
- 8/ The notation is standard: $o_s(n^\alpha)$ denotes a matrix whose elements are the random variables a_{ijn} each with $\lim_{n \rightarrow \infty} a_{ijn}/n^\alpha = 0$ almost surely, $O_s(n^\alpha)$ denotes a matrix whose elements are the random variable a_{ijn} each with a_{ijn}/n^α is bounded almost surely. Similarly $o_p(n^\alpha)$ and $O_p(n^\alpha)$ for convergence in probability.

REFERENCES

- [1] Barnett, W. A. (1976). Maximum likelihood and iterated Aitken estimation in nonlinear systems of equations. J. Amer. Statist. Assoc. 71 354-360.
- [2] Caves, D. W. and Christensen, L. R. (1978). Global properties of of flexible functional forms. Technical report. Social Systems Research Institute, University of Wisconsin.
- [3] Christensen, L. R., Jorgenson, D. W., and Lau, L. J. (1975). Transcendental logarithmic utility functions. Amer. Econ. Rev. 65 367-383.
- [4] Chung, K. L. (1974). A Course in Probability, 2nd ed. Academic Press.
- [5] Gallant, A. R. (1975). Seemingly unrelated nonlinear regressions. J. Econometrics 3 35-50.

- [6] Gallant, A. R. (1977). Three-stage least squares estimation for a system of simultaneous, nonlinear, implicit equations. J. Econometrics 5 71-88.
- [7] Gallant, A. R. and Holly, A. (1978). Statistical inference for an implicit, nonlinear, simultaneous equation model in the context of maximum likelihood estimation. Cahiers du Laboratoire d'Econometrie A 191 1078, Ecole Polytechnique, Paris.
- [8] Holly, A. (1978). Tests of nonlinear statistical hypotheses in multiple equation nonlinear models. Cahiers du Laboratoire d'Econometrie A 176 0178, Ecole Polytechnique, Paris.
- [9] Huber, P. J. (1964). The behavior of maximum likelihood estimators under nonstandard conditions. Proc. Fifth Berkeley Symp. Math. Statist. Prob. 1 73-101.
- [10] Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares. Ann. Math. Statist. 40 633-643.
- [11] Malinvaud, E. (1970). The consistency of nonlinear regressions. Ann. Math. Statist. 41 965-969.
- [12] Searle, S. R. (1971). Linear Models. Wiley.