STATISTICAL INFERENCE IN NONLINEAR MODELS:
A PSEUDO LIKELIHOOD APPROACH

by

GERALDO DA SILVA E SOUZA

A thesis submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

DEPARTMENT OF STATISTICS

RALEIGH
1979

APPROVED BY:

[Signatures]

Chairman of Advisory Committee
Geraldo da Silva e Souza was born September 19, 1946, in Rio de Janeiro, Brasil. He received his elementary and secondary education in his home town, graduating from Colegio Estadual Prof. Clovis Monteiro in 1968. He entered Universidade Federal do Rio de Janeiro, Rio de Janeiro, and received a Bachelor of Science degree in Mathematics in 1971. He also entered Universidade do Estado do Rio de Janeiro, Rio de Janeiro, and received a Bachelor of Science degree in Economics in 1972. In January, 1973, the author was granted a fellowship from the Conselho Nacional de Pesquisas, Rio de Janeiro, to pursue graduate studies at Universidade Federal do Rio de Janeiro. He received a Master of Science degree in System Engineering from this university in 1974.

In August, 1974, the author accepted employment with Empresa Brasileira de Correios e Telegrafos, Brasilia, as an economist. In December, 1974, he was named Chairman of the Division of Economic Studies. In July, 1975, he accepted a part-time employment with the Department of Statistics at Universidade de Brasilia, Brasilia, as an Assistant Professor. In August, 1976, he was granted financial support from Empresa Brasileira de Pesquisa Agropecuaria, Brasilia, and enrolled in the Department of Statistics at North Carolina State University to pursue a program of studies leading to the Doctoral degree in Statistics.

The author is married to the former Wanda Fritsch. Mrs. Souza graduated from Escola Nacional de Ciencias Estatisticas, Rio de Janeiro, and from Universidade do Estado do Rio de Janeiro, Rio de Janeiro, in 1972, receiving the degree of Bachelor of Science in Statistics and Bachelor of Science in Economics, respectively. They have two children: Alexandre, born February 9, 1973; and Marcos, born October 1, 1974.
1. INTRODUCTION .............................................. 1
2. LITERATURE REVIEW ........................................ 2
3. THE STATISTICAL MODEL ................................. 7
   3.1 Description ........................................... 7
   3.2 Probability Space ................................. 8
   3.3 Estimation .......................................... 9
   3.4 Hypothesis Testing .............................. 10
4. CESARO SUMMABILITY ................................. 11
5. CONSISTENCY ............................................. 23
6. ASYMPTOTIC NORMALITY ............................... 29
7. TEST STATISTICS .............................. 40
   7.1 Wald's Test ....................................... 40
   7.2 The Lagrange Multiplier Test .................... 42
   7.3 The Likelihood Ratio Test ...................... 44
   7.4 Applications ..................................... 47
8. THE MAXIMUM LIKELIHOOD ESTIMATOR .............. 50
9. THE SINGLE EQUATION EXPLICIT MODEL .......... 53
10. THE SEEMINGLY UNRELATED NONLINEAR MODEL .... 57
11. LIST OF REFERENCES ............................... 62
1. INTRODUCTION

The general regression context is that of an observed multivariate response to an observed multivariate input. Based on empirical evidence or some underlying theory the investigator decides on the kind of structure (model) linking input variables to response variables. Frequently, the nature of the problem under study gives rise to a statistical model with additive errors

\[ G(y_t, x_t, \gamma) = e_t. \]

This representation defines responses (y's), implicitly, as a function of exogenous variables (inputs, the x's), unknown parameters (\gamma), and random components (e's). This thesis considers estimation and hypothesis testing for such models. The approach adopted is the inclusion, in the model's description, of an optimization criterion which defines an estimator. This is called the pseudo likelihood. A theory of large sample statistical inference is then developed for the pseudo likelihood estimator including its strong consistency, asymptotic normality, and the null and non null distributions of the Wald's test statistic, an analog of the likelihood ratio test statistic, and an analog of the Lagrange multiplier test statistic (Rao's efficient score test). It turns out that the proposed theory encompasses a variety of statistical procedures in use for nonlinear models. In this class we may include Zellner-type nonlinear multivariate procedures, and the more standard least squares and maximum likelihood procedures.
2. LITERATURE REVIEW

Jenrich (1969) and Malinvaud (1970) were the precursors of what is known as large sample statistical theory of nonlinear regression models. Jenrich's paper, dealing with the single equation explicit model

\[ y_t = g(x_t, \gamma) + e_t \quad t = 1, 2, \ldots, n \]

sets forth conditions for the strong consistency and asymptotic normality of the least squares estimator and shows that the Gauss–Newton iteration method of estimation is asymptotically stable. Based on assumptions of uniform convergence Jenrich establishes a method of proving strong consistency, and a method of proving asymptotic normality without requiring third order derivatives.

Malinvaud's paper develops a description of the requisite behavior of exogenous variables in terms of weak convergence of measures so as to achieve the uniform convergence required by Jenrich. Malinvaud applies his methods both to single equation explicit models and to multivariate regressions.

After the papers of Jenrich and Malinvaud the field developed rapidly. A variety of models have been considered and rigorously studied. In the context of univariate explicit models, Gallant (1973) combines the techniques of Malinvaud and Jenrich to obtain strong consistency and asymptotic normality of the nonlinear least squares estimator without assuming the existence of second order derivatives of the response function. In the discussion of the regression problem compactness of the parameter space is not required. In another series of papers Gallant (1975a, 1975b, 1977b) develops a complete theory of inference
for all practical purposes including Monte Carlo simulations. These papers together with Blattburg, McGuire, and Weber (1971) reveal that the convergence (in distribution) of the likelihood ratio test statistic is faster than that of the Wald's test statistic. Still in the context of the single equation explicit model a good summary of results may be found in Gallant (1975d). This article is an expository presentation of the theory and methods of nonlinear regression which exploits the analogies with the theory and methods of linear regression.

In the context of multivariate explicit models, i.e., models of the form

\[ y_{\alpha t} = g_{\alpha}(x_{\alpha t}, \gamma_{\alpha}) + e_{\alpha t} \quad t = 1, \ldots, n; \alpha = 1, 2, \ldots, m \]

Gallant (1975c) proves strong consistency and asymptotic normality of the three step procedure estimator. Barnett (1976) working with the same model approaches the regression problem by means of maximum likelihood estimation.

Single equation explicit models with time series errors have been studied by Robinson (1972), Hannan (1971), and Gallant and Goebel (1976). These models are of the form

\[ y_t = g(x_t, \gamma) + u_t \quad t = 1, \ldots, n \]

where the \( u_t \)'s are assumed to define a covariance stationary time series. The main approach to these models has been through an initial approximation of the variance covariance matrix of the \( u \)'s by means of nonlinear least squares residuals and subsequent use of the Aitken procedure.
Ballet Lawrence (1975), dealing with the single equation implicit model

\[ g(y_t, x_t, \gamma) = e_t \quad t = 1, \ldots, n \]

introduces a distance function-absolute value of \( g \) to the \( p \)th power \( (p \geq 1) \), to define an estimator. Strong consistency is proved for \( p \geq 1 \). Asymptotic normality is obtained for \( p > 1 \). An analytic solution for \( y \) is not required in applications. Asymptotic properties of a predictor are also studied. Her work is a precursor of many of the ideas which appear in this thesis.

The more general case of simultaneous systems of nonlinear, implicit equations, have been the most recent area of development. Amemya (1974) proves weak consistency and asymptotic normality of the two stage nonlinear least squares. Jorgenson and Lafont (1974) derive a Cramer-Rao lower bound and show that maximum likelihood theory can be used to generate an estimator that attains this bound. Amemya (1977) develops the maximum likelihood approach. He proves weak consistency and asymptotic normality of the maximum likelihood estimator based on an assumption of normality for the error distribution. He also proves that the maximum likelihood estimator is more efficient than the nonlinear three stage least squares estimator if the model specification is correct. Gallant (1977a) describes the nonlinear three stage least squares estimator. Strong consistency and asymptotic normality is proved. His methods allow parametric estimation subject to nonlinear restrictions across equations. Gallant and Holly (1978) refine the concept of Cesaro summability - a uniform strong law of large numbers, previously introduced in
Gallant (1977a). They obtain, without normality assumptions, the asymptotic law of maximum likelihood estimators (which is normal). A complete large sample theory of statistical inference is developed in this paper.

This thesis is concerned with methods and ideas introduced by Gallant and Holly (1978). Thus it is worthwhile to describe in more detail the framework upon which their results are based. The residual sequence is generated according to the density \( p(e|\sigma^*) \), where \( \sigma^* \) is an unknown parameter. The model, in implicit form, is

\[
g(y_t, x_t, \theta^0) = e_t \quad t = 1, 2, \ldots, n.
\]

The conditions of the implicit function theorem are satisfied so that the model induces a conditional density function for \( y \) of the form

\[
p(y|x, \theta^0, \sigma^*) .
\]

The only structure exploited is that of an estimator \( \hat{\gamma} = (\hat{\theta}, \hat{\sigma}) \) defined by maximizing

\[
\frac{1}{n} \sum_{t=1}^{n} \ln p(y_t|x_t, \theta, \sigma) .
\]

Assuming that

\[
\iint \ln p(y|x, \theta, \sigma)p(y|x, \theta^0, \sigma^*)dy \, P_X(dx)
\]

has a unique maximum over the parameter space at \( (\theta, \sigma) = (\theta^0, \sigma^*) \), where \( P_X \) is a convenient measure on \( X \), strong consistency of \( \hat{\gamma} \) can be proved. The limiting normal law of \( \hat{\gamma} \) is obtained by means of the
asymptotic normality of the scores

\[ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{3}{\partial \gamma} \ln p(y_t | x_t, \theta^0, \sigma^*) \ ; \gamma = (\theta, \sigma) . \]

Here we follow precisely along these lines. Instead of the likelihood a general function is considered and an optimization procedure is defined. The properties of the resulting estimator are studied based on the unique maximum assumption. A score is defined and its asymptotic law is obtained. The resulting theory covers not only the Gallant and Holly (1978) results but also most of the methods and models described above. Thus, this thesis contributes to a unification of this literature.
3. THE STATISTICAL MODEL

3.1 Description

The basic model is a relation of the form

\[ G(y_t, x_t, \gamma^0) = e_t \quad t = 1, 2, \ldots, n. \]

The \( y \)'s are sample values of endogenous variables, vectors in \( Y \), a Borel subset of \( \mathbb{R}^m \). The \( x \)'s are sample values of exogenous variables, vectors in \( X \), a Borel subset of \( \mathbb{R}^k \). The \( e \)'s are sample values of residual variables, vectors in \( E \), a Borel subset of \( \mathbb{R}^m \). The true parameter \( \gamma^0 \) is a vector in \( \Gamma \), a subset of \( \mathbb{R}^p \).

Assumption 1

For each \( (x, \gamma) \in X \times \Gamma \) the structural equation

\[ G(y, x, \gamma) = e \]

defines a one to one mapping from \( Y \) onto \( E \). This implies the existence of a reduced form \( y = f(e, x, \gamma) \) which maps \( E \) onto \( Y \) and satisfies

\[ G(f(e, x, \gamma), x, \gamma) = e. \]

The function \( f \) defining the reduced form is continuous on \( E \times X \times \Gamma \).

We remark that in any application of the model proposed here, in order to use the statistical methods we will set forth, it will not be necessary to have a closed form expression for the reduced form, or even to be able to compute it using numerical methods.
Assumption 2

The residual sequence \((e_t)_{t \geq 1}\) is a realization of a sequence of identically and independently distributed random vectors. Their common probability measure, defined on the Borel subsets of \(E\), is denoted by \(P_E\).

We emphasize that the distribution determined by \(P_E\) is parameter free. In many instances the case envisaged is when residual variables have mean zero and identity variance covariance matrix. In applications, if any parametrization is present in the error distribution the model should be transformed so that all parameters are absorbed in \(\gamma\). The models we study in Chapters 8, 9, and 10 are typical examples of this case.

Assumption 3

The set \(X\) is provided with a probability measure \(P_X\). The sequence of exogenous variables may be fixed or random. No matter how exogenous variables are treated (random or not), the sequence \((x_t)_{t \geq 1}\) is held fixed.

For the moment we only want the reader to note that \((x_t)_{t \geq 1}\) will be treated as a fixed sequence in a probability space. A formal link between \(P_X\) and \((x_t)_{t \geq 1}\) will be established in Chapter 4.

3.2 Probability Space

Let Assumptions 2 and 3 hold. Let \(E\) denote the set of all sequences of elements in \(E\). By the infinite dimensional product measure theorem (Ash (1972)), \(P_E\) determines a probability measure \(P\) on the class \(\mathcal{B}(E)\) of the Borel subsets of \(E\). The probability space \((E, \mathcal{B}(E), P)\) provides the basic probability space for the discussion of
our statistical problems. Thus, in what follows, probability statements will refer to \( P \) and the random variables, for the particular sequence \((x_t)_{t \geq 1}\) fixed, will be functions defined on \( E \) and measurable with respect to \( \mathcal{B}(E) \).

### 3.3 Estimation

For estimation purposes interest is focused on an \( \lambda \)-vector of parameters \( \lambda^0 \). In a typical application, \( \lambda^0 \) will equal some easily computed function of \( \gamma^0 \). The vector \( \lambda^0 \) is a point in \( \Lambda \).

The estimation process will be carried out with the use of a real valued function \( s(y,x,\tau,\lambda) \) defined on \( Y \times X \times T \times \Lambda \) where \( T \) is a subset of \( \mathbb{R}^q \). The variable \( \tau \) corresponds to the presence of a parameter \( \tau^* \in T \). In a typical application, \( \tau^* \) will equal some easily computed function of \( \gamma^0 \) and will be regarded as a nuisance parameter.

**Assumption 4**

The function \( s(y,x,\tau,\lambda) \) is continuous. A strongly consistent estimator \( \hat{\tau}_n \) is available for estimating \( \tau^* \). The sequence \( (\sqrt{n} (\hat{\tau}_n - \tau^*))_{n \geq 1} \) is bounded in probability, i.e., given \( \varepsilon > 0 \) there exists a bound \( M_\varepsilon \) such that for \( n \) sufficiently large,

\[
P(\sqrt{n} |\hat{\tau}_n - \tau^*| < M_\varepsilon) > 1 - \varepsilon.
\]

**Definition 3.3.1**

Let Assumptions 1 through 4 hold. Any sequence of random variables \( (\hat{\lambda}_n)_{n \geq 1} \) satisfying

\[
S_n(\hat{\lambda}_n) = \sup_{\lambda \in \Lambda} S_n(\lambda)
\]

satisfies...
where

$$S_n(\lambda) = \frac{1}{n} \sum_{t=1}^{n} s(y_t, \hat{x}_t, \hat{y}_n, \lambda)$$

is called a sequence of pseudo likelihood estimators for $\lambda^0$. We refer to the function $s$ as the pseudo likelihood.

In Chapter 5 we show that under certain conditions there exists a sequence of pseudo likelihood estimators strongly consistent for $\lambda^0$.

The formulation considered here for 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. In Chapters 8, 9, and 10 we present examples which suggest possible uses for the pseudo likelihood and the unifying character of the theory we are to develop.

3.4 Hypothesis Testing

In this thesis we address the problem of testing the hypothesis $H_0: h(\lambda^0) = 0$ against the alternative $H_A: h(\lambda^0) \neq 0$, where $h$ maps $\Lambda$ into $\mathbb{R}^r$. Test statistics will be defined in Chapter 7. The theorems on the asymptotic distributions of those random variables will permit not only the statistical test of $H_0$ against $H_A$ but also the evaluation of the power of the corresponding test for local alternatives.
4. CESARO SUMMABILITY

In this chapter we discuss the notion of Cesaro summable sequences. The concept refines Malinvaud's (1970) use of the weak convergence of measures in nonlinear statistical analysis, and has been used by Gallant (1977a) and Gallant and Holly (1978) to describe the requisite asymptotic behavior of the joint sequence \((e_t, x_t)_{t \geq 1}\). The main consequences of its use here are:

i) A conclusion similar to that of the Helly-Bray theorem is feasible without the restrictive boundedness condition.

ii) Regularity conditions that insure the existence of the almost sure uniform limit of sequences of the form

\[
\frac{1}{n} \sum_{t=1}^{n} \phi(f(e_t, x_t, \gamma), x_t, \rho)
\]

can be stated in terms of properties that are intrinsic to the statistical model.

Definition 4.1

A sequence of points \((v_t)_{t \geq 1}\) in a Borel subset \(V\) of a Euclidean space is said to generate Cesaro summable sequences with respect to the probability measure \(\nu\), defined on the Borel subsets of \(V\), and the function \(b(v)\), nonnegative, \(\nu\)-integrable, if

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \alpha(v_t) = \int \alpha(v) \nu(dv)
\]

for every function \(\alpha(v)\), real valued, continuous, and dominated by \(b(v)\).
The notion of Cesaro summability is not void. We present an elementary proof of this fact here using Stute's (1976) generalization of the Glivenko-Cantelli theorem. This result, as pointed out by Wesler (1979), can be obtained with weaker hypotheses using more advanced methods of proof.

**Theorem 4.1**

Let \((Z^t)_{t=1}^\infty\) be a sequence of identically and independently distributed M-dimensional random variables defined on the probability space \((\Omega,\mathcal{A},P^*)\). Let \(\nu\) be the common induced probability measure defined on the Borel subsets of \(\mathbb{R}^M\). Assume \(\nu\) absolutely continuous with respect to some product measure on \(\mathbb{R}^M\). Let \(b(\nu)\) be nonnegative and \(\nu\)-integrable. There exists a set \(A \in \mathcal{A}\) such that \(P^*(A) = 0\) and if \(w \notin A\)

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \alpha(Z^t(w)) = \int_{\mathbb{R}^M} \alpha(\nu) \nu(d\nu)
\]

for every function \(\alpha(\nu)\), real valued, continuous, and dominated by \(b(\nu)\).

**Proof:**

Let \(V\) be the collection of all bounded cubes in \(\mathbb{R}^M\). By Stute's (1976) generalization of the Glivenko-Cantelli theorem there exists a \(P^*\)-null set \(A_0 \subset A\) such that if \(w \notin A_0\)

\[
\limsup_{n \to \infty} \sup_{V \in V} \left| \frac{1}{n} \sum_{t=1}^{n} I_V(Z^t(w)) - \nu(V) \right| = 0 \tag{4.1}
\]

where \(I_V\) is the notation for the indicator function of a set \(V\).
Since \( b(v) \) is an integrable function one may choose a sequence \( V_j \) of cubes such that

\[
\int_{V_j} b(v) \nu (dv) < \frac{1}{j}.
\]

By the strong law of large numbers for each cube \( V_j \) there exists a set \( A_j \) with \( P^*(A_j) = 0 \) such that for every \( w \in A_j^c \) there exists a positive integer \( n_j \) with the property that

\[
\frac{1}{n} \sum_{t=1}^{n} I_{V_j}^{c}(Z_t(w)) b(Z_t(w)) < \frac{1}{j}; \quad n \geq n_j.
\]

Let \( A = \bigcup_{j=0}^{\infty} A_j \) and \( \varepsilon > 0 \) be given. Let \( w \in A^c \) and denote \( z_t = Z_t(w) \). There exists \( n_j \) and \( V_j \) such that

\[
\int_{V_j} b(v) \nu (dv) < \frac{\varepsilon}{5},
\]

\[
\frac{1}{n} \sum_{t=1}^{n} I_{V_j}^{c}(z_t) b(z_t) < \frac{\varepsilon}{5} \quad \text{for} \quad n \geq n_j.
\]

Let \( V_o = V_{j_o} \) and \( a(v) \) continuous and dominated by \( b(v) \). It is possible to divide \( V_o \) into disjoint cubes \( V_i \); \( i = 1, 2, \ldots, i_o \), and construct \( g \) of the form

\[
g(v) = \sum_{i=1}^{i_o} r_i I_{V_i}(v); \quad |r_i| > 0; \quad i = 1, 2, \ldots, i_o.
\]
such that

$$\sup_{v \in V_0} |a(v) - g(v)| < \frac{\varepsilon}{5}. \quad (4.3)$$

By (4.1) there exists \( n_\alpha > n_j \) such that for \( n \geq n_\alpha \)

$$\sup_{v \in V} \left| \frac{1}{n} \sum_{t=1}^{n} I_v(z_t) - \nu(v) \right| < \frac{\varepsilon}{5} \sum_{i=1}^{n} |r_i| \quad (4.4)$$

If \( n > n_\alpha \), then

$$\left| \frac{1}{n} \sum_{t=1}^{n} a(z_t) - \int a(v) \nu(dv) \right|$$

$$\leq \left| \frac{1}{n} \sum_{t=1}^{n} I_{V_0}(z_t) a(z_t) - \int a(v) \nu(dv) \right| + \frac{1}{n} \sum_{t=1}^{n} I_{V_0}(z_t) b(z_t) + \int b(v) \nu(dv)$$

$$\leq \left| \frac{1}{n} \sum_{t=1}^{n} I_{V_0}(z_t) a(z_t) - \int a(v) \nu(dv) \right| + \frac{2\varepsilon}{5} \text{ by } (4.2).$$

Also,

$$\left| \frac{1}{n} \sum_{t=1}^{n} I_{V_0}(z_t) a(z_t) - \int a(v) \nu(dv) \right|$$

$$\leq \frac{1}{n} \sum_{t=1}^{n} I_{V_0}(z_t) |a(z_t) - g(z_t)| + \left| \frac{1}{n} \sum_{t=1}^{n} I_{V_0}(z_t) g(z_t) - \int g(v) \nu(dv) \right|$$

$$+ \int |a(v) - g(v)| \nu(dv)$$
< \left| \frac{1}{n} \sum_{t=1}^{n} I_{\mathcal{V}_0}(z_t) g(z_t) - \int g(v) \nu(dv) \right| + \frac{2\varepsilon}{5} \quad \text{by (4.3).}

Now

\left| \frac{1}{n} \sum_{t=1}^{n} I_{\mathcal{V}_0}(z_t) g(z_t) - \int g(v) \nu(dv) \right|

= \left| \sum_{i=1}^{i_0} r_i \left[ \frac{1}{n} \sum_{t=1}^{n} I_{\mathcal{V}_i}(z_t) - \nu(\mathcal{V}_i) \right] \right|

\leq \sum_{i=1}^{i_0} |r_i| \quad \varepsilon = \frac{\varepsilon}{5} \quad \text{by (4.4).}

\sum_{i=1}^{i_0} |r_i| < \sum_{i=1}^{i_0} \varepsilon = \frac{\varepsilon}{5}

Hence

\left| \frac{1}{n} \sum_{t=1}^{n} a(z_t) - \int a(v) \nu(dv) \right| < \varepsilon \quad \text{for } n \geq n_\alpha \quad \Box

Let Assumptions 2 and 3 hold. We now motivate a crucial assumption regarding Cesaro summability and the model introduced in Chapter 3 by means of two examples which follow directly from Theorem 4.1. The first example shows that the property of Cesaro summability holds for the joint sequence \((e_t, x_t)_{t \geq 1}\) (and also for the components) when exogenous variables are fixed in repeated samples, \(P_E\) is absolutely continuous, and \(P_X\) is chosen to be the limiting measure introduced by Malinvaud (1970). The second example deals with the random case, i.e., when exogenous variables are random. It shows that the same conclusion holds if independence and absolute continuity are imposed.
Example 1:

Let $M$ be a positive integer and $a_0, \ldots, a_{M-1}$ a finite collection of vectors in $\mathbb{R}^k$. Assume that $(x_t)_{t \geq 0}$ is generated according to

$$x_t = a_t \mod M$$

and that $P_E$ is absolutely continuous with respect to some product measure on $\mathbb{R}^m$. Then $(x_t)_{t \geq 0}$ generates Cesaro summable sequences with respect to

$$P_X(A) = \frac{1}{M} \sum_{t=0}^{M-1} I_A(a_t).$$

The dominating function in this case is irrelevant (in the sense that domination is not necessary) as well as the continuity requirement. Moreover, if $\mu = P_E \times P_X$ and $b(e, x)$ is nonnegative and $\mu$-integrable, almost every sequence $(e_t)_{t \geq 1}$ is such that $(e_t, x_t)_{t \geq 1}$ generates Cesaro summable sequences with respect to $\mu$ and $b(e, x)$.

Indeed, let $\alpha(x)$ be a measurable function on $X$ and $n > M$. We can write $n = aM + c$, where $0 \leq c \leq M - 1$, and

$$\frac{1}{n} \sum_{t=0}^{n} \alpha(x_t) \leq \frac{1}{n} \left[ \sum_{t=0}^{M-1} \alpha(x_t) + \sum_{t=M}^{2M-1} \alpha(x_t) + \ldots + \sum_{t=(a-1)M}^{aM-1} \alpha(x_t) + \sum_{t=aM}^{aM+c} \alpha(x_t) \right]$$

$$= \frac{a}{aM+c} \sum_{t=0}^{M-1} \alpha(a_t) + \frac{1}{aM+c} \sum_{t=M}^{aM+c} \alpha(a_t)$$
Thus,

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{t=0}^{n} \alpha(x_t) = \frac{1}{M} \sum_{t=0}^{M-1} \alpha(a_t) = \int_{X} \alpha(x) P_X(dx).
\]

On the other hand let \( \phi(e,x) \) be continuous on \( E \times X \) and dominated by \( b(e,x) \). For \( n = aM + c, 0 \leq c < M - 1 \), we can write

\[
\frac{1}{n} \sum_{t=0}^{n} \phi(e_t, x_t) = \frac{1}{n} \left[ \sum_{t=0}^{M-1} \phi(e_t, x_t) + \sum_{t=M}^{2M-1} \phi(e_t, x_t) + \ldots + \sum_{t=(a-1)M}^{aM-1} \phi(e_t, x_t) \right]
\]

\[
+ \sum_{t=aM}^{aM+c} \phi(e_t, x_t)
\]

\[
= \frac{1}{n} \left[ \sum_{i=0}^{c} \sum_{j=0}^{a} \phi(e_{Mj+i}, a_i) + \sum_{i=c+1}^{M-1} \sum_{j=0}^{a-1} \phi(e_{Mj+i}, a_i) \right]
\]

\[
= \frac{a+1}{aM+c} \sum_{i=0}^{c} \sum_{j=0}^{a} \phi(e_{Mj+i}, a_i) + \frac{a}{aM+c} \sum_{i=c+1}^{M-1} \sum_{j=0}^{a-1} \phi(e_{Mj+i}, a_i).
\]

By theorem 4.1 there exist null sets \( A(a_0), \ldots, A(a_{M-1}) \) such that if \( (e_t)_{t \geq 1} \) is not a realization in \( \bigcup_{j=0}^{M-1} A(a_i) \), then

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{t=0}^{n} \phi(e_t, x_t) = \frac{1}{M} \sum_{i=0}^{M-1} \frac{1}{M} \int_{E} \phi(e, a_i) P_{E}(de)
\]

\[
= \int_{E} \int_{X} \phi(e, x) \mu(de, dx),
\]

by Fubini's theorem (Ash(1972)). \( \square \)
Example 2:

Assume that the sequence of exogenous variables is independent of the sequence of residual variables and defines an independently and identically distributed stochastic process with common distribution \( P_X \). Assume also that \( P_X \) and \( P_E \) are absolutely continuous with respect to product measures on \( \mathbb{R}^k \) and \( \mathbb{R}^m \), respectively. Let \( X \) denote the set of all sequences of elements in \( X \). A sequence in \( X \) will be denoted by \( \bar{x} \) and a sequence in \( E \) by \( \bar{e} \). Let \( P' \) be the probability measure determined by \( P_X \) on the Borel subsets of \( X \) and \( \nu \) the product probability on \( E \times X \) generated by \( P \) and \( P' \). Let \( b(e,x) \) be a measurable function on \( E \times X \), nonnegative, \( \mu \)-integrable, where \( \mu = P_E \times P_X \). Then the following conditions hold:

1) There exists a \( P' \)-null set \( B \) such that every realization \( \bar{x} \) in \( B^c \) generates Cesaro summable sequences with respect to \( P_X \) and

\[
b(x) = \int_E b(e,x) P_E(de) .
\]

ii) For every realization \( \bar{x} \) in \( B^c \) there exists a \( P \)-null set \( A(\bar{x}) \) in \( B(E) \) such that for every \( \bar{e} \) in \( [A(\bar{x})]^c \), \( (\bar{e}, \bar{x}) \) generates Cesaro summable sequences with respect to \( \mu \) and \( b(e,x) \).

Indeed, by theorem 4.1 there exists a \( \nu \)-null set \( A \) such that Cesaro summability with respect to \( \mu \) and \( b(e,x) \) holds for every joint sequence \( (\bar{e}, \bar{x}) \) in \( A^c \). For each \( \bar{x} \in X \) consider the measurable section \( A(\bar{x}) = \{ \bar{e} \in E; (\bar{e}, \bar{x}) \in A \} \). Since

\[
\nu(A) = \int_X P(A(\bar{x})) P'(d\bar{x}) = 0
\]
there exists a \( P' \)-null set \( B_1 \) such that if \( \bar{x} \notin B_1 \), \( P(A(\bar{x})) = 0 \)
and on \([A(\bar{x})]^c\) every sequence \( \bar{e} \) will be such that \((\bar{e}, \bar{x})\) generates
Cesaro summable sequences with respect to \( \mu \) and \( b(e, x) \). By theorem
4.1 there exists a \( P' \)-null set \( B_2 \) such that on \( B_2^c \) every sequence
\( \bar{x} \) generates Cesaro summable sequences with respect to \( P_X \) and \( b(x) \).
Take \( B = B_1 \cup B_2 \). \( \Box \)

Thus, it is plausible to impose

Assumption 5

Let \( \mu \) be the product probability on the Borel sets of \( E \times X \)
determined by \( P_E \) and \( P_X \) (Assumptions 2 and 3). There exists a
function \( b(e, x) \), measurable, nonnegative, \( \mu \)-integrable such that

i) For every \( x_0 \in X \) there exists a neighborhood \( N(x_0) \) such
that \( \sup_{x \in N(x_0)} b(e, x) \) is measurable and \( P_E \)-integrable.

ii) The sequence \((x_t)_{t \geq 1} \) (Assumption 3) generates Cesaro
summable sequences with respect to \( P_X \) and

\[
b(x) = \int_E b(e, x) P_E(de).
\]

iii) There exists a set \( A \) (depending on \((x_t)_{t \geq 1}\)) in
\( E(E) \) such that \( P(A) = 0 \) and for every sequence \((e_t)_{t \geq 1} \) in \( A^c \),
\((e_t, x_t)_{t \geq 1} \) generates Cesaro summable sequences with respect to \( \mu \)
and \( b(e, x) \).

Condition (i) of Assumption 5 is a technical requisite that will
imply continuity of certain functions defined through processes of
integration. Examples 1 and 2 show that conditions (ii) and (iii) are
verified in instances that come readily to mind.
The proof of the following lemma is in Gallant (1977a).

Lemma 4.1

Let $a(v, \theta)$ be real valued, continuous, defined on $V \times \Theta$, where $V$ is a Borel set and $\Theta$ is compact ($V$ and $\Theta$ are both subsets of Euclidean spaces). Let $(v_t)_{t \geq 1}$ be a sequence in $V$ generating Cesaro summable sequences with respect to the probability measure $\nu$ on $V$ and a dominating function $b(v)$. If

$$\sup_{\theta \in \Theta} |a(v, \theta)| \leq b(v)$$

for every $v$ in $V$, then:

i) $\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} a(v_t, \theta) - \int_{V} a(v, \theta) \nu(dv) = 0$.

ii) $\int_{V} a(v, \theta) \nu(dv)$ is continuous on $\Theta$.

Theorem 4.2

Let Assumptions 1, 2, 3, and 5 hold. Let $\Gamma'$ be a compact subset of $\Gamma$. Let $\phi(y, x, \rho)$ be real valued and continuous on $Y \times X \times K$, where $K$ is a compact subset of a Euclidean space. If

$$|\phi(y, x, \rho)| \leq b(G(y, x, \gamma), x)$$

for all $(y, x, \gamma, \rho)$ in $Y \times X \times \Gamma' \times K$, then for almost every realization $(e_t)_{t \geq 1}$
\[ \frac{1}{n} \sum_{t=1}^{n} \phi(f(e_t, x_t, \gamma), x_t, \rho), \]

\[ \frac{1}{n} \sum_{t=1}^{n} \int_{E} \phi(f(e, x_t, \gamma), x_t, \rho) \, P_{E}(de), \]

converge uniformly on \( \Gamma' \times K \) to

\[ \int_{E} \int_{X} \phi(f(e, x, \gamma), x, \rho) \mu(de, dx). \]

The limit function is a continuous function of \((\gamma, \rho)\).

**Proof:**

Given \((e, x, \gamma, \rho)\), let \(y = f(e, x, \gamma)\). Then

\[ |\phi(f(e, x, \gamma), x, \rho)| = |\phi(y, x, \rho)| \leq b(G(y, x, \gamma), x) = b(e, x). \]

Hence

\[ \sup_{(\gamma, \rho) \in \Gamma' \times K} |\phi(f(e, x, \gamma), x, \rho)| \leq b(e, x), \]

\[ \sup_{(\gamma, \rho) \in \Gamma' \times K} \left| \int_{E} \phi(f(e, x, \gamma), x, \rho) \, P_{E}(de) \right| \leq b(x). \]

By condition (i) of Assumption 5 and the dominated convergence theorem
\[
\int_{E} \phi(f(e, x, y), x, \rho) \, P_{E}(de)
\]

is continuous. The conclusion then follows from Lemma 4.1. □
5. CONSISTENCY

In this chapter we establish conditions under which we may guarantee the existence and strong consistency of a sequence of pseudo likelihood estimators. In order to use our results on estimation in the derivation of the asymptotic non-null distribution of test statistics it is necessary to introduce a dependence of $\gamma^O$ (therefore of $\lambda^O$) on the sample size.

Assumption 6

The true parameter $\gamma^O$ is indexed by $n$ and $\lim_{n \to \infty} \gamma^O_n = \gamma^*$, $\gamma^* \in \Gamma$. There are unique points $\lambda = \lambda^*, \lambda^O_1, \lambda^O_2, \ldots$ corresponding to $\gamma = \gamma^*, \gamma^O_1, \gamma^O_2, \ldots$, which maximize

$$S(\gamma, \tau^*, \lambda) = \int \int s(f(e, x, \gamma), x, \tau^*, \lambda) \mu(de, dx)$$

over $\Lambda$. The function $f$ is the reduced form of Assumption 1 and $\mu$ is the probability measure of Assumption 5. Also $\lim_{n \to \infty} \sqrt{n} (\lambda^O_n - \lambda^*) = \delta$, $\delta \in \mathbb{R}^2$.

Notice that if Assumption 1 holds and we are given a sequence $(\gamma^O_n)_{n \geq 1}$ in $\Gamma$ and a joint sequence $(e_t, x_t)_{t \geq 1}$ we may use the reduced form $f(e, x, \gamma)$ to define a triangular array in $Y$ as follows. For each $n \geq 1$, let $\gamma_{tn} = f(e_t, x_t, \gamma^O_n)$ for $1 \leq t \leq n$. This is indeed the structure we have in mind when Assumption 6 is present. In this context the sequence $(\gamma_t)_{t \geq 1}$ defined in Chapter 3 may be viewed as a particular case. It is the array corresponding to the constant sequence $\gamma^O_n = \gamma^O$. In this case Assumption 6 also defines the
conceptual link between $\lambda^0$ and $\gamma^0$. The vector $\lambda^0$, on which we focus our interest for estimation purposes, must be the unique point producing the maximum of $S(y^0, \tau^*, \lambda)$ over $\Lambda$. As we are going to see only a parameter satisfying this condition can be estimated consistently. This is a conclusion from theorem 5.1 below which states that $\hat{\lambda}_n$ (definition 3.3.1) converges almost surely to $\lambda^*$ when Assumptions 1 through 6 (and other technical conditions) hold. Thus, any result proved in the context of Assumption 6 may be translated in terms of the model proposed in Chapter 3.

Another point to be explained is the presence of $\tau^*$ when Assumption 6 holds. In that general setting, $\tau^*$ is to be viewed as a nuisance parameter defined as the almost sure limit of $\hat{\tau}_n$.

From this point on, whenever Assumption 6 is present, we will use the notation $y_{\tau n}$ suggested above. In this context $S_n(\lambda)$ of Definition 3.3.1, for example, becomes

$$S_n(\lambda) = \frac{1}{n} \sum_{t=1}^{n} s(y_{\tau n}, x_t, \tau_n, \lambda).$$

**Definition 5.1**

Let $\epsilon > 0$ be given and $\theta$ a point in a Euclidean space. The open ball with center at $\theta$ and radius $\epsilon$ is denoted $N_{\epsilon}(\theta)$. The corresponding closed ball is represented by $\overline{N}_{\epsilon}(\theta)$.

**Assumption 7**

The set $\Lambda$ is closed and unbounded. There exists a positive function $w(\lambda)$ for which:
i) $\sup_{\lambda \in \Lambda} \frac{s(y, x, I, \lambda)}{w(\lambda)}$ is continuous.

ii) There exists $\varepsilon_0 > 0$ such that

$$\sup_{\lambda \in \Lambda} \left| \frac{s(y, x, I, \lambda)}{w(\lambda)} \right| \leq b(G(y, x, \gamma), x)$$

for all $(y, x, \gamma, I)$ in $Y \times X \times \Pi_{\varepsilon_0} (\gamma^*) \times \Pi_{\varepsilon_0} (I^*)$, where

$\Pi_{\varepsilon_0} (\gamma^*) \subseteq \Gamma$ and $\Pi_{\varepsilon_0} (I^*) \subseteq T$.

iii) $\liminf_{||\lambda|| \to \infty} w(\lambda) > -d$, where $d = S(\gamma^*, I^*, \lambda^*)$.

The conditions stated in Assumption 7 represent a modification of those presented by Huber (1967) and are necessary in the proof of Lemma 5.1 below which establishes the existence of pseudo likelihood estimators. If compactness of $\Lambda$ is assumed, Lemma 5.1 is not necessary and domination is all that is needed to prove strong consistency.

**Lemma 5.1**

Let Assumptions 1 through 7 hold. There is a compact set $\Lambda'$ containing $\lambda^*$ such that for almost every realization $(e_t)_{t \geq 1}$ there exists a positive integer $n_o$ for which $n \geq n_o$ implies

$$\sup_{\lambda \in \Lambda} S_n(\lambda) = \sup_{\lambda \in \Lambda', n \geq n_o} S_n(\lambda).$$

**Proof:**

By condition (iii) of Assumption 7 there exists a compact set $\Lambda'$ and $0 < \varepsilon < 1$ such that

$$\inf_{\lambda \in \Lambda'} w(\lambda) \geq \frac{-d + \varepsilon}{1 - \varepsilon}.$$
Let \((e_t^*)_{t>1}\) be a realization in \(A^c\) (Assumption 5) such that 
\((\hat{\tau}_n^*)_{n>1}\) is convergent to \(\tau^*\). Almost every error realization is such.

By conditions (i), (ii), and (iii) of Assumption 7 and Theorem 4.2, there exists \(n_1\) such that for \(n > n_1\)

\[
\inf_{\lambda \in \Lambda'} \frac{1}{n} \sum_{t=1}^{n} \frac{s(y_{tn}, x_t, \hat{\tau}_n, \lambda)}{w(\lambda)} \geq \frac{1}{n} \sum_{t=1}^{n} \inf_{\lambda \in \Lambda} \frac{s(y_{tn}, x_t, \hat{\tau}_n, \lambda)}{w(\lambda)} > 1 - \varepsilon.
\]

Hence, if \(\lambda \notin \Lambda'\) and \(n > n_1\)

\[
\frac{1}{n} \sum_{t=1}^{n} s(y_{tn}, x_t, \hat{\tau}_n, \lambda) < d - \varepsilon.
\]

By condition (ii) of Assumption 7, and Theorem 4.2

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \frac{s(y_{tn}, \hat{\tau}_n^*, \lambda^*)}{w(\lambda^*)} = \frac{d}{w(\lambda^*)}
\]

Thus, there exists \(n_2\) such that

\[
\frac{1}{n} \sum_{t=1}^{n} s(y_{tn}, x_t, \hat{\tau}_n, \lambda^*) > d - \varepsilon
\]

for \(n > n_2\). Hence \(\lambda^* \in \Lambda'\). Let \(n_0 = \max\{n_1, n_2\}\). If \(n \geq n_0\) the maximum of \(S_n(\lambda)\) must occur at some point of \(\Lambda'\).

Lemma 5.1 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 Assumption 7, one may verify this conclusion directly as in Gallant and Holly (1978).
Let $\hat{\lambda}_n$ be such that $S_n(\hat{\lambda}_n) = \sup_{\lambda \in \Lambda'} S_n(\lambda)$. Appealing to Lemma 2 of Jenrich (1969), stated below, we may consider $\hat{\lambda}_n$ a random variable.

**Lemma 5.2**

Let $Q$ be a real valued function on $\Theta \times Y$ where $\Theta$ is a compact set of a Euclidean space and $Y$ is a measurable space. For each $\theta$ in $\Theta$ let $Q(\theta, y)$ be a measurable function of $y$ and for each $y$ in $Y$ a continuous function of $\theta$. Then there exists a measurable function $\hat{\theta}$ from $Y$ into $\Theta$ such that for all $y$ in $Y$

$$Q(\hat{\theta}(y), y) = \inf_{\Theta \in \Theta} Q(\theta, y).$$

**Theorem 5.1 (Strong consistency)**

Let Assumptions 1 through 7 hold. Let $(\hat{\lambda}_n)_{n \geq 1}$ be a sequence of random variables (which exists by Lemmas 5.1 and 5.2) such that for $n$ sufficiently large $S_n(\hat{\lambda}_n) = \sup_{\lambda \in \Lambda'} S_n(\lambda)$. Then $\lim_{n \to \infty} \hat{\lambda}_n = \lambda^*$ almost surely.

**Proof:**

Let $(e_t)_{t \geq 1}$ be a realization in $A^c$ (Assumption 5) such that $(\hat{\tau}_n)_{n \geq 1}$ is convergent. Almost every error realization is such.

By condition (ii) of Assumption 7 and Theorem 4.2 $(S_n(\lambda))_{n \geq 1}$ converges uniformly on $\Lambda'$ to $S(\gamma^*, \tau^*, \lambda)$. Let $\lambda$ be a limit point of $(\hat{\lambda}_n)_{n \geq 1}$. There exists a subsequence $(\hat{\lambda}_{n_k})_{k \geq 1}$ converging to $\lambda$. Then

$$S(\gamma^*, \tau^*, \lambda) = \lim_{n_k \to \infty} S_n(\hat{\lambda}_{n_k}) \geq \lim_{n_k \to \infty} S_n(\lambda^*) = S(\gamma^*, \tau^*, \lambda^*).$$
By Assumption 6 \( S(\gamma^*, \tau^* \lambda) \) has a unique maximum at \( \lambda = \lambda^* \). Hence
\[
\bar{\lambda} = \lambda^*.
\]
Since for \( n \) large \( \hat{\lambda}_n \) is confined to a compact set,
\[
\lim_{n \to \infty} \hat{\lambda}_n = \lambda^*.
\]
6. ASYMPTOTIC NORMALITY

Frequently we will make use of the matrix representation of certain concepts in differential calculus. The notation employed is as follows. Let \( \phi(x,y) \) be a real valued function of the \( k \)-vector \( x \) and the \( m \)-vector \( y \). We denote

\[
\frac{\partial \phi}{\partial x} (x,y) = \left( \frac{\partial \phi}{\partial x_1} (x,y), \ldots, \frac{\partial \phi}{\partial x_k} (x,y) \right)'
\]

\[
\frac{\partial \phi}{\partial x'} (x,y) = \left( \frac{\partial \phi}{\partial x} (x,y) \right)'
\]

If \( W(x) = (\phi_1(x), \ldots, \phi_c(x))' \) is a vector valued function of \( x \)

\[
\frac{\partial W}{\partial x'} (x) = \begin{bmatrix}
\frac{\partial \phi_1}{\partial x'} (x) \\
\vdots \\
\frac{\partial \phi_c}{\partial x'} (x)
\end{bmatrix}
\]

Also,

\[
\frac{\partial^2 \phi}{\partial y \partial x_1} (x,y) = \frac{\partial}{\partial y'} \left( \frac{\partial \phi}{\partial x_1} (x,y) \right),
\]

\[
\frac{\partial^2 \phi}{\partial y' \partial x} (x,y) = \frac{\partial}{\partial y'} \left( \frac{\partial \phi}{\partial x} (x,y) \right).
\]
In particular,

\[ \frac{\partial^2 \phi}{\partial x \partial y} (x,y) = \frac{\partial}{\partial x} \left[ \frac{\partial \phi}{\partial x} (x,y) \right] \]

The basic rules for manipulation of derivatives given in matrix form may be found in Theil (1971).

Let Assumptions 1 through 7 hold so that there exists a sequence of random variables \( \hat{\lambda}_n \) strongly consistent for \( \lambda^* \) satisfying \( S_n (\hat{\lambda}_n) = \sup_{\lambda \in \Lambda} S_n (\lambda) \) for \( n \) sufficiently large. The framework necessary to establish asymptotic normality for \( \sqrt{n} (\hat{\lambda}_n - \lambda^*) \) is determined by Assumption 8 below. The notation is the same considered in Assumptions 1 through 7.

**Assumption 8**

The elements of \( \frac{\partial S}{\partial \lambda} (y,x,\tau,\lambda), \frac{\partial^2 S}{\partial \lambda^2} (y,x,\tau,\lambda), \frac{\partial^2 S}{\partial \tau \partial \lambda} (x,y,\tau,\lambda), \) and \( \frac{\partial S}{\partial \tau} (y,x,\tau,\lambda) \) are continuous and dominated by \( b(G(y,x,\gamma),x) \) for all \( (y,x,\gamma,\tau,\lambda) \) in \( Y \times X \times N_{e_0} (\gamma^*) \times N_{e_0} (\tau^*) \times N_{e_0} (\lambda^*) \), where \( N_{e_0} (\lambda^*) \leq \Lambda' \). Moreover

i) For every \( x \in X \) and \( n \geq 1 \)

\[ \int \frac{\partial S}{\partial \lambda} (f(e,x,\gamma_n^0),x,\tau^*,\lambda_n^0) P_E (de) = 0. \]

ii) \( \int \int \frac{\partial^2 S}{\partial \tau \partial \lambda} (f(e,x,\gamma^*),x,\tau^*,\lambda^*) \mu (de, dx) = 0. \)

The two integral conditions imposed in Assumption 8 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. Our example in Chapter 10 serves to illustrate both conditions.

Definition 6.1

\[
J = - \int \int \frac{\partial^2 S}{\partial \lambda' \partial \lambda} (f(e, x, y^*), x, \tau^*, \lambda^*) \mu(dx, dy)
\]

\[
I = \int \int \frac{\partial S}{\partial \lambda} (f(e, x, y^*), x, \tau^*, \lambda^*) \frac{\partial S}{\partial \tau} (f(e, x, y^*), x, \tau^*, \lambda^*) \mu(dx, dy)
\]

Proposition 6.1

Let Assumptions 1 through 8 hold. Then \( J \) is positive definite.

Proof:

By the mean value theorem and dominated convergence theorem

\[
J = - \frac{\partial^2 S}{\partial \lambda' \partial \lambda} (y^*, \tau^*, \lambda^*)
\]

where \( S \) is defined in Assumption 6. Since \( S(y^*, \tau^*, \lambda) \) has a unique maximum at \( \lambda^* \), \( -J \) must be negative definite. Then \( J \) must be positive definite. \( \square \)

At this point we remark that at the level of generality considered here the matrix \( I \) (information matrix) may not be equal to \( J \) or positive definite. These facts do not affect our proofs of asymptotic normality in Theorems 6.1 and 6.2. However, the singularity of \( I \) is a nuisance for hypothesis testing and does not seem to be of practical importance. In Chapter 7 we will impose nonsingularity for the information matrix. As shown in Gallant and Holly (1978), \( I \) is
non-singular (indeed $I = J$) when one restricts attention to maximum likelihood estimation.

Lemma 6.1 below is proved in Jenrich (1968). It is a version of the mean value theorem from advanced calculus for random functions. The proof presented by Jenrich permits an immediate extension to Taylor's expansions.

**Lemma 6.1**

Let $\phi(y, \theta)$ be a real valued function on $Y \times \Theta$ where $Y$ is a measurable space and $\Theta$ is a convex compact subset of a Euclidean space. For each $\theta$ in $\Theta$ let $\phi(y, \theta)$ be a measurable function of $y$ and for each $y$ in $Y$ a continuously differentiable function of $\theta$. Let $\theta_1$ and $\theta_2$ be measurable functions from $Y$ into $\Theta$. Then there exists a measurable function $\bar{\theta}$ from $Y$ into $\Theta$ such that

(i) $\phi(y, \theta_1(y)) - \phi(y, \theta_2(y)) = \frac{\partial \phi}{\partial \theta} (y, \bar{\theta}(y)) (\theta_1(y) - \theta_2(y))$.

(ii) $\bar{\theta}(y)$ lies on the segment joining $\theta_1(y)$ and $\theta_2(y)$ for all $y$ in $Y$.

We now introduce the analog of the scores defined by Gallant and Holly (1978).

**Theorem 6.1**

Let Assumptions 1 through 8 hold. Then

$$\sqrt{n} \frac{\partial \bar{\theta}}{\partial \lambda} (\lambda^0) \xrightarrow{L} N(0, I)$$
Proof:

The proof is done in two steps. First we show that

$$
\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial S}{\partial \lambda} \left( y_{tn}, x_t^*, \tau_n^*, \lambda_n^o \right) \xrightarrow{L} N(0, I)
$$

verifying the Lindeberg-Feller condition. Notice that this random variable differs from $$\sqrt{n} \frac{\partial S}{\partial \lambda} \left( \lambda_n^o \right)$$ only in the $$\tau$$ component. The latter has $$\hat{\tau}_n$$ instead of $$\tau^*$$. With this result on hand then we show that the conclusion of the theorem is implied by Assumption 4 and the second integral condition of Assumption 8.

We may assume without any loss of generality that $$\hat{\tau}_n$$ takes its values in $$N_{\varepsilon o} (\tau^*)$$ for every $$n$$ (it is possible to construct a tail equivalent sequence with this property).

Let $$u$$ be a vector with $$|u| = 1$$ and consider

$$
Z_{tn} (e_t) = u' \frac{\partial S}{\partial \lambda} (f(e_t, x_t^*, y_n^o, x_t^*, \tau_n^*, \lambda_n^o) \ 1 \leq t \leq n \ ; \ n = 1, 2, \ldots .
$$

Each $$Z_{tn}$$ has mean zero by the integral condition (i) of Assumption 8, and variance

$$
\sigma^2_{tn} = u' \int \frac{\partial S}{\partial \lambda} (f(e, x_t^*, y_n^o, x_t^*, \tau_n^*, \lambda_n^o) \frac{\partial S}{\partial \tau} (f(e, x_t^*, y_n^o, x_t^*, \tau_n^*, \lambda_n^o) P_E (de) u .
$$

By Theorem 4.2,

$$
\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \sigma^2_{tn} = u' I u . \quad (6.1)
$$

If $$u' I u = 0$$, $$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} Z_{tn}$$ converges in distribution to zero by Chebychev's inequality. Assume $$u' I u > 0$$. Let $$\varepsilon > 0$$ be given and
$$A_n = \{ z \in \mathbb{R} ; \ |z| > \varepsilon \sqrt{V_n} \} \ ; \ V_n = \sum_{t=1}^{n} \sigma_{tn}^2 .$$

Let

$$B_n = \frac{1}{n} \sum_{t=1}^{n} \int_{\mathbb{R}} I_{A_n}(Z_{tn}(e)) Z_{tn}^2(e) P_{\mathbb{E}}(de) .$$

By the transformation theorem (Ash (1972))

$$\frac{1}{V_n} \sum_{t=1}^{n} \int_{A_n} Z_{tn}^2 \ dF_{Z_{tn}} = \frac{n}{V_n} B_n .$$

We now show that the Lindeberg-Feller condition for Chung's (1974) Central Limit Theorem is satisfied, i.e., $$\lim_{n \to \infty} \frac{n}{V_n} B_n = 0 .$$ By (6.1) $$\lim_{n \to \infty} \frac{n}{V_n} = (u'u)^{-1} > 0 .$$ Thus it is enough to show that $$\lim_{n \to \infty} B_n = 0 .$$ For this purpose let $$\rho = (\gamma, \lambda), \rho_n = (\gamma_n, \lambda_n^0), \rho^* = (\gamma^*, \lambda^*),$$ and for $$a \geq 0$$

$$A(a) = \{ z \in \mathbb{R} ; \ |z| > sa \} .$$

Let

$$\mathcal{B}_a(\rho) = \int_{E} \int_{X} I_{A(a)}(u' \frac{3s}{\partial \lambda} (f(e,x,\gamma^*),x,\tau^*,\lambda^*)) x (u' \frac{3s}{\partial \lambda} (f(e,x,\gamma^*),x,\tau^*,\lambda^*)) u(de,dx) .$$

Since $$\mathcal{B}_o(\rho^*)$$ is finite (Assumption 8), given $$\eta > 0$$ there exists $$a_o$$ such that $$\mathcal{B}_a(\rho^*) < \frac{\eta}{2} .$$
Choose a continuous function \( \phi \) on the line and a positive integer \( n_1 \) such that

\[
I_{a_{n_1}}(z) \leq \phi(z) \leq I_{a_{a_{n_1}}}(z)
\]

for every \( n \geq n_1 \).

Let

\[
B_\phi(\rho) = \frac{1}{n} \sum_{t=1}^{n} \int_E \phi(u, \frac{\partial s}{\partial \lambda} (f(e, x_t, \gamma), x_t, \tau^*, \lambda)) \times (u, \frac{\partial s}{\partial \lambda} (f(e, x_t, \gamma), x_t, \tau^*, \lambda))^2 p_E(\mathrm{d}e).
\]

By Theorem 4.2, \( \lim_{n \to \infty} B_\phi(\rho_n) = B_\phi(\rho^*) \). Hence there exists a positive integer \( n_2 \) such that for all \( n \geq n_2 \)

\[
B_\phi(\rho_n) < B_\phi(\rho^*) + \frac{n}{2}.
\]

Let \( n > \max(n_1, n_2) \). Then

\[
B_n \leq B_\phi(\rho_n) < B_\phi(\rho^*) + \frac{n}{2} < n.
\]

Thus, \( \lim_{n \to \infty} B_n = 0 \) and

\[
\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial s}{\partial \lambda} (y_{tn}, x_t, \tau^*, \lambda^c_n) \xrightarrow{L} N(0, I).
\]

(6.2)
Let

$$g_n(\tau) = \frac{1}{n} \sum_{t=1}^{n} \frac{\partial S}{\partial \lambda}(y_{tn}, x_{tn}\tau, \lambda_n^0)$$

and denote by $g_{ni}(\tau)$ the $i$-th component of $g_n(\tau)$. By Lemma 6.1, for each $i = 1, 2, \ldots, l$ there exists a sequence of random variables $(\tau_{ni})_{n\geq 1}$ converging to $\tau^*$ for which

$$g_{ni}(\tau_n) - g_{ni}(\tau^*) = \frac{\partial g_{ni}}{\partial \tau}(\tau_{ni})(\tau_n - \tau^*)$$

By Theorem 4.2 and the integral condition (ii) of Assumption 8,

$$\lim_{n \to \infty} \frac{\partial g_{ni}}{\partial \tau}(\tau_{ni}) = 0$$

almost surely. Then we can write

$$g_n(\tau_n) - g_n(\tau^*) = C_n(\tau_n - \tau^*)$$

where the elements of $C_n$ go to zero almost surely. Since $\sqrt{n}(\tau_n - \tau^*)$ is bounded in probability (Assumption 4), $\sqrt{n}g_n(\tau_n)$ and $\sqrt{n}g_n(\tau^*)$ have the same asymptotic distribution. But

$$\sqrt{n}g_n(\tau_n) = \sqrt{n} \frac{\partial S}{\partial \lambda}(\lambda_n^0),$$

$$\sqrt{n}g_n(\tau^*) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial S}{\partial \lambda}(y_{tn}, x_{tn}\tau^*, \lambda_n^0).$$

The result then follows from (6.2).
Theorem 6.2 (Asymptotic normality)

Let Assumptions 1 through 6 hold. Then

\[ \sqrt{n} \frac{3S_n}{\partial \lambda} (\lambda^*) \xrightarrow{L} N(J_\theta I), \]

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

Let

\[ \hat{J}_n = - \frac{1}{n} \sum_{t=1}^{n} \frac{\partial^2 S}{\partial \lambda^t \partial \lambda}(y_t, x_t^t, \hat{\tau}_n, \hat{\lambda}_n), \]

\[ \hat{I}_n = \frac{1}{n} \sum_{t=1}^{n} \frac{\partial S}{\partial \lambda}(y_t, x_t^t, \hat{\tau}_n, \hat{\lambda}_n) \frac{\partial^2 S}{\partial \lambda}(y_t, x_t^t, \hat{\tau}_n, \hat{\lambda}_n). \]

Then, almost surely, \( \lim_{n \to \infty} \hat{J}_n = J \) and \( \lim_{n \to \infty} \hat{I}_n = I. \)

Proof:

Without any loss of generality we may assume that \( \hat{\lambda}_n \) takes its values in \( \bar{N}_{\theta_0} (\lambda^*) \) and \( \hat{\tau}_n \) in \( \bar{N}_{\theta_0} (\tau^*) \), for every \( n \geq 1 \) (it is possible to construct tail equivalent sequences with this property).

The convergence of \( \hat{J}_n \) and \( \hat{I}_n \) follows from Theorem 4.2.

By Lemma 6.1, for each \( i = 1, 2, \ldots, k \) there exists a sequence of random variables \( (\bar{\lambda}_{ni})_{n \geq 1} \) converging to \( \lambda^* \) for which

\[ \frac{\partial S_n}{\partial \lambda_i} (\hat{\lambda}_n) - \frac{\partial S_n}{\partial \lambda_i} (\lambda^*) = \frac{\partial^2 S_n}{\partial \lambda_i \partial \lambda_i} (\bar{\lambda}_{ni}) (\hat{\lambda}_n - \lambda^*). \]
Since $\hat{\lambda}_n$ maximizes $S_n(\lambda)$ for $n$ large, $\frac{\partial S_n}{\partial \lambda}(\hat{\lambda}_n) = 0$. Thus

$$\frac{\partial S_n}{\partial \lambda} (\lambda^*) = -\frac{\partial^2 S_n}{\partial \lambda^2}(\hat{\lambda}_n)(\hat{\lambda}_n - \lambda^*)$$

and by Theorem 4.2 we can write

$$\frac{\partial S_n}{\partial \lambda} (\lambda^*) = (J^+ a_n)(\hat{\lambda}_n - \lambda^*) \quad (6.3)$$

where the elements of $a_n$ go to zero almost surely. A similar argument yields

$$\frac{\partial S_n}{\partial \lambda} (\lambda^*) - \frac{\partial S_n}{\partial \lambda} (\lambda^0_n) = (J^+ \beta_n)(\lambda^0_n - \lambda^*) \quad (6.4)$$

where the elements of $\beta_n$ go to zero.

By (6.4), Theorem 6.1, and Assumption 6

$$\sqrt{n} \frac{\partial S_n}{\partial \lambda} (\lambda^*) \xrightarrow{L} N(J\delta, I) .$$

From this fact and (6.3)

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

We translate these results to the language of Chapter 3 as follows. Suppose the vector $\lambda^0$ satisfies the unique maximum condition. Then, under the assumptions we have stated so far, $\hat{\lambda}_n$ is strongly consistent
for $\lambda^0$ and $\sqrt{n} (\hat{\lambda}_n - \lambda^0) \xrightarrow{L} N(0, J^{-1}J - 1)$. The asymptotic variance covariance matrix is estimated consistently by $J^{-1}\hat{J}_n^{-1}$. 
7. TEST STATISTICS

Tests of the hypothesis $H_0$: $h(\lambda^o) = 0$ against $H_A$: $h(\lambda^o) \neq 0$ are considered here. Test statistics are defined and the asymptotic distributions of these random variables are derived in the context of Assumptions 1 through 8 and certain conditions imposed on $h$. These results, as seen in Section 7.4, dictate the procedures to be followed in any application of the statistical model proposed in Chapter 3.

Assumption 9

The matrix $I$ of Definition 6.1 has full rank.

Assumption 10

The r-vector valued function $h$ is continuously differentiable, $h(\lambda^*) = 0$, and $H(\lambda^*) = \frac{\partial h}{\partial \lambda^T}(\lambda^*)$ has full row rank. The statement "under the null hypothesis" means that $h(\lambda^o_n) = 0$ for every $n$.

In what follows we use Searle's (1971) definition of a noncentral chi-square.

7.1 Wald's test

The Wald's test statistic is

$$W_n = n(h(\hat{\lambda}_n))(\hat{H}_n \hat{\nu}^{-1} \hat{H}_n)^{-1}(h(\hat{\lambda}_n))$$

where

$$\hat{H}_n = H(\hat{\lambda}_n) = \frac{\partial h}{\partial \lambda^T}(\hat{\lambda}_n),$$

$$\hat{\nu}_n = \hat{j}_T^{-1} n \hat{j}_n^{-1}.$$
Theorem 7.1.1

Let Assumptions 1 through 10 hold. Then $W_n$ converges in distribution to a noncentral chi-square with $r$ degrees of freedom and noncentrality parameter $\frac{\delta' \{H'(HVH')^{-1}H\delta\}}{2}$, where $H = H(\lambda^*)$, and $V = J^{-1}IJ^{-1}$. Under the null hypothesis the convergence is to a chi-square with $r$ degrees of freedom.

Proof:

Without any loss of generality we may assume that $\hat{\lambda}_n$ takes its values in $\tilde{N}_c(\lambda^*)$ for every $n$ (we can construct a tail equivalent sequence with this property). By Lemma 6.1, for each $i = 1, 2, \ldots, r$ there exists a sequence of random variables $(\tilde{\lambda}_{ni})_{n \geq 1}$ converging to $\lambda^*$ almost surely for which

$$h_i(\hat{\lambda}_n) - h_i(\lambda^*) = \frac{\partial h_i}{\partial \lambda^i}(\tilde{\lambda}_{ni})(\hat{\lambda}_n - \lambda^*) .$$

Since the elements of $H(\lambda)$ are continuous functions we may write

$$h(\hat{\lambda}_n) - h(\lambda^*) = (H + \alpha_n)(\hat{\lambda}_n - \lambda^*)$$

where the elements of $\alpha_n$ go to zero almost surely. It follows that $\sqrt{n} h(\hat{\lambda}_n)$ and $H\sqrt{n} (\hat{\lambda}_n - \lambda^*)$ have the same asymptotic distribution.

Since $(\hat{H}_n \hat{u}_n \hat{H}_n)'^{-1/2}$ exists for $n$ sufficiently large and converges almost surely to $(HVH')^{-1/2}$,

$$(\hat{H}_n \hat{u}_n \hat{H}_n)'^{-1/2} \sqrt{n} h(\hat{\lambda}_n) ,$$
and

\[(HVH')^{-\frac{1}{2}} \sqrt{n} (\lambda_n - \lambda^*)\]

have the same asymptotic distribution. Since

\[(HVH')^{-\frac{1}{2}} \sqrt{n} (\lambda_n - \lambda^*) \xrightarrow{L} \mathcal{N}(HVH')^{-\frac{1}{2}}H\delta, I_r), \tag{7.1}\]

the first claim of the theorem follows.

Assume \(h(\lambda_o^n) = 0\) for every \(n\). By the mean value theorem, for each \(i = 1, 2, \ldots, r\) there exists a sequence \((\lambda_{ni})_{n \geq 1}\) converging to \(\lambda^*\) for which

\[h_i(\lambda^o_n) - h_i(\lambda^*) = \frac{\partial h_i}{\partial \lambda^i}(\lambda_{ni}^o)(\lambda_{ni} - \lambda^*).\]

Thus we can write

\[(H + \beta_n) \sqrt{n} (\lambda^o_n - \lambda^*) = 0\]

where the elements of \(\beta_n\) are convergent to zero. It follows that \(H\delta = 0\). The convergence of \(W_n\) to a chi-square with \(r\) degrees of freedom follows from (7.1). \(\square\)

### 7.2 The Lagrange Multiplier Test

The Lagrange multiplier test (an analog of it) requires the existence of an estimator subject to the constraint \(h(\lambda) = 0\). If Assumptions 1 through 10 hold there exists a sequence of random variables \((\lambda_n)_{n \geq 1}\) such that for \(n\) sufficiently large \(\lambda_n\) maximizes \(S_n(\lambda)\)
subject to $h(\lambda) = 0$. An argument similar to that of Theorem 5.1 shows that except for outcomes in an unusual set, $\lim_{n \to \infty} \tilde{\lambda}_n = \lambda^*$. The Lagrange multiplier test statistic is

$$R_n = n \frac{3S}{3\lambda} (\tilde{n})^{-1} J_n^{-1} H_n (H_n^{-1} V_n H_n)^{-1} J_n^{-1} \frac{3S}{3\lambda} (\tilde{n})$$

where $\tilde{J}_n$, $\tilde{H}_n$, and $\tilde{V}_n$ are the analogs of $\hat{J}_n$, $\hat{H}_n$, and $\hat{V}_n$, respectively, when the constrained estimator $\tilde{\lambda}_n$ is considered.

**Theorem 7.2.1**

Let Assumptions 1 through 10 hold. Then $R_n$ converges in distribution to a noncentral chi-square with $r$ degrees of freedom and noncentrality parameter $\frac{\delta' H'(HVH')^{-1} H \delta}{2}$. Under the null hypothesis, $R_n$ converges in distribution to a chi-square with $r$ degrees of freedom.

**Proof:**

Without any loss of generality we may assume that $\tilde{\lambda}_n$ takes its values in $N_{\infty} (\lambda^*)$ for every $n$ (we can construct a tail equivalent sequence with this property).

By Lemma 6.1, Theorem 4.2, and almost sure convergence of $\hat{\lambda}_n$ and $\hat{\tau}_n$, we can write

$$\frac{3S}{3\lambda} (\lambda^*) - \frac{3S}{3\lambda} (\tilde{\lambda}_n) = (J + \alpha_n)(\tilde{\lambda}_n - \lambda^*)$$

where the elements of $\alpha_n$ go to zero almost surely.

By Lemma 6.1, continuity of the elements of $H(\lambda)$, and almost sure convergence of $\tilde{\lambda}_n$, we can write
\[ h(\hat{\lambda}_n^*) - h(\lambda^*) = (H + \beta_n)(\hat{\lambda}_n^* - \lambda^*) \]

where the elements of \( \beta_n \) go to zero almost surely. Hence, for \( n \) large,

\[ \sqrt{n} (H + \beta_n) (J + \alpha_n)^{-1} \left( \frac{\partial S_n}{\partial \lambda} (\lambda^*) - \frac{\partial S_n}{\partial \lambda} (\hat{\lambda}_n^*) \right) = \sqrt{n} (H + \beta_n) (\hat{\lambda}_n^* - \lambda^*) = 0 . \]

Thus

\[ \sqrt{n} \, H J^{-1} \, \frac{\partial S_n}{\partial \lambda} (\hat{\lambda}_n^*) \xrightarrow{L} N(H\delta, HVH') . \quad (7.2) \]

Since \( \hat{J}_n, \hat{V}_n, \) and \( \hat{H}_n \) converge almost surely to \( J, V, \) and \( H, \) respectively,

\[ (H_n \, V_n \, H_n)^{-1} H_n \, J_n^{-1} \, \sqrt{n} \, \frac{\partial S_n}{\partial \lambda} (\hat{\lambda}_n^*) \xrightarrow{L} N[(HVH')^{-1/2} H\delta, I_r] \]

and the first claim of the theorem follows. Under the null hypothesis, as in the proof of theorem 7.1.1, \( H\delta = 0 . \)

### 7.3 The Likelihood Ratio Test

The likelihood ratio test (an analog of it) is based on the statistic

\[ L_n = 2n(\lambda_n^* - S_n(\hat{\lambda}_n^*) - S_n(\lambda_n^*)) . \]
Theorem 7.3.1

Let Assumptions 1 through 10 hold. Assume also \( J = I \). Then \( L_n \) converges in distribution to a noncentral chi-square with \( r \) degrees of freedom and noncentrality parameter \( \frac{\delta' H' (HVH')^{-1} H \delta}{2} \). Under the null hypothesis, \( L_n \) converges in distribution to a chi-square with \( r \) degrees of freedom.

Proof:

Without any loss of generality we may assume that \( \lambda_n \) and \( \lambda_n^* \) take their values in \( \overline{N}_{\epsilon_0} (\lambda^*) \) (we can construct tail equivalent sequences with this property).

By the Taylor's expansion version of Lemma 6.1 there exists a sequence of random variables \( (\lambda_n^*)_{n \geq 1} \) converging to \( \lambda^* \) almost surely for which

\[
S_n(\lambda_n) = S_n(\lambda_n^*) + \frac{\partial S_n}{\partial \lambda} (\lambda_n^*)(\lambda_n^* - \lambda_n^*) + \frac{1}{2} \frac{\partial^2 S_n}{\partial \lambda \partial \lambda'} (\lambda_n^*)(\lambda_n^* - \lambda_n^*) (\lambda_n^* - \lambda_n^*).
\]

By Theorem 4.2 we can write

\[
2n(S_n(\lambda_n) - S_n(\lambda_n^*)) = n(\lambda_n - \lambda_n^*)' (J + \alpha_n)(\lambda_n - \lambda_n^*)
\]

(7.3)

where the elements of \( \alpha_n \) go to zero almost surely.

There are Lagrange multipliers \( \theta_n \) satisfying

\[
\frac{\partial S_n}{\partial \lambda} (\lambda_n^*) - H_n \theta_n = 0.
\]
By continuity of the elements of \( H(\lambda) \) and almost sure convergence of \( \lambda_n \), we can write \( H_n = H + \beta_n \), where the elements of \( \beta_n \) go to zero almost surely. Then

\[
\sqrt{n} \left[ H J^{-1} \frac{\partial S_n}{\partial \lambda} (\lambda_n) - H J^{-1} (H + \beta_n)' \sqrt{n} \theta_n \right] = 0.
\]

From this expression and (7.2) in the proof of Theorem 7.2.1, we conclude

\[
\sqrt{n} \theta_n \overset{L}{\rightarrow} N[(H J^{-1} H', H J^{-1} H')^{-1}(H V H') (H J^{-1} H')]. \tag{7.4}
\]

By Lemma 6.1, Theorem 4.2, and almost sure convergence of \( \lambda_n, \lambda_n', \) and \( \tau_n' \), we can write

\[
\frac{\partial S_n}{\partial \lambda} (\lambda_n) - \frac{\partial S_n}{\partial \lambda} (\lambda_n') = (J + \xi_n)(\lambda_n - \lambda_n')
\]

where the elements of \( \xi_n \) go to zero almost surely. Then, for \( n \) large,

\[
(H + \alpha_n)' \sqrt{n} \theta_n = (J + \xi_n)' \sqrt{n} (\lambda_n - \lambda_n')
\]

since \( \frac{\partial S_n}{\partial \lambda} (\lambda_n) = 0 \) by definition of \( \lambda_n \). Thus,

\[
(J + \xi_n)^{-1} (H + \alpha_n)' \sqrt{n} \theta_n = \sqrt{n} (\lambda_n - \lambda_n')
\]
and by (7.4)

\[ \sqrt{n}(\hat{\lambda}_n - \tilde{\lambda}_n) \xrightarrow{L} N[J^{-1}H'(HJ^{-1}H')^{-1}H\delta, J^{-1}H'(HJ^{-1}H')^{-1}(HVH')(HJ^{-1}H')^{-1}HJ^{-1}] \]  

(7.5)

Hence

\[ 2n[S_n(\hat{\lambda}_n) - S_n(\tilde{\lambda}_n)] \xrightarrow{L} L(Z'JZ) \]

where \( Z \) is the multivariate normal with mean and variance covariance matrix as in (7.5). If \( J = I \) then \( V = J^{-1} \) and the variance covariance matrix of \( Z \), \( \Sigma \) say, is such that \( \Sigma J \) is idempotent.

The first claim of the theorem then follows from Searle (1971). Under the null hypothesis, as before, \( H\delta = 0 \). □

### 7.4 Applications

As we have done in previous chapters we now show how the theory fits in the statistical model proposed in Chapter 3. If \( \lambda^0 \) satisfies the unique maximum requirement, we have seen that our assumptions imply the almost sure convergence of \( \hat{\lambda}_n \) to \( \lambda^0 \) and asymptotic normality for \( \sqrt{n}(\hat{\lambda}_n - \lambda^0) \). Suppose we want to perform a test of \( H_0 \) against \( H_A \).

For large samples, under \( H_0 \), the statistics \( W_n, R_n \), and \( L_n \) (if \( I = J \)) have distributions that can be approximated by a chi-square with \( r \) degrees of freedom and therefore may be used for the test. Suppose we also want to investigate the behavior of the particular test statistic we choose in the presence of small deviations from the null hypothesis. Our theory says that a noncentrality parameter is obtained if we specify a sequence (convergent) \( \gamma_n \) such that the corresponding
sequence \((\lambda_n^0)_{n \geq 1}\) (Assumption 6) converges to a point satisfying the null hypothesis. To use this result in practice the argument is as follows. As mentioned earlier, typically, \(\lambda\) and \(\tau\) are easily computed functions of \(\gamma\), so that if \(\gamma_n^0\) is specified then \(\lambda_n^0\) and \(\tau_n^0\) becomes specified and this \(\lambda_n^0\) will satisfy the unique maximum condition. Thus, in a typical application, \((\gamma_n^0, \tau_n^0, \lambda_n^0)\) is specified and the noncentrality parameter \(\alpha = \frac{\delta' H'(HV'H')^{-1}H\delta}{2}\) is to be computed.

The use of Theorem 7.4.1 below avoids the problem of having to specify \((\gamma^*, \tau^*, \lambda^*)\) to make this computation. The interesting feature of this result is that integration with respect to \(P_x\) is not necessary to approximate \(\alpha\).

**Theorem 7.4.1**

Let Assumptions 1 through 10 hold. Let \((\gamma_n^0, \tau_n^0, \lambda_n^0)\) be any sequence converging to \((\gamma^*, \tau^*, \lambda^*)\) and \(\lim_{n \to \infty} \sqrt{n}(\lambda_n^0 - \lambda^*) = 0\). Let

\[
I_n = \frac{1}{n} \sum_{t=1}^{n} \int_E \frac{\partial^2 s}{\partial \lambda^2} \left( f(e, x, \gamma_n^0, x_t, \tau_n^0, \lambda_n^0) - f(e, x, \gamma_n^0, x_t, \tau_n^0, \lambda_n^0) \right) P_x(de),
\]

\[
J_n = -\frac{1}{n} \sum_{t=1}^{n} \int_E \frac{\partial^2 s}{\partial \lambda^2} \left( f(e, x, \gamma_n^0, x_t, \tau_n^0, \lambda_n^0) \right) P_x(de),
\]

\[
H_n^0 = H(\lambda_n^0),
\]

\[
\alpha_n^0 = n(h(\lambda_n^0))'(H_n^0 V_n^0 H_n^0)^{-1}(h(\lambda_n^0)).
\]

Then \(\lim_{n \to \infty} \alpha_n^0 = \alpha\).
Proof:

By the mean value theorem, for each \( i = 1, \ldots, r \) there exists a sequence \((\overline{\lambda}_n^i)_{n \geq 1}\) converging to \( \lambda^* \) for which

\[
h_i(\lambda_n^0) - h_i(\lambda^*) = \frac{3h_i}{3\lambda} (\overline{\lambda}_n^i)(\lambda_n^0 - \lambda^*).
\]

By continuity of the elements of \( H(\lambda) \) we can write

\[
h(\lambda_n^0) = (H + \beta_n)(\lambda_n^0 - \lambda^*)
\]

where the elements of \( \beta_n \) converge to zero. Since \((H_n^0 \Sigma_n^0 H_n^0)^{-1}\) exists for \( n \) sufficiently large, \((H_n^0 \Sigma_n^0 H_n^0)^{-1/2}\) exists and

\[
(H_n^0 \Sigma_n^0 H_n^0)^{-1/2} \sqrt{n} h(\lambda_n^0) = (H_n^0 \Sigma_n^0 H_n^0)^{-1/2} (H + \beta_n) \sqrt{n} (\lambda_n^0 - \lambda^*)
\]

Hence

\[
\lim_{n \to \infty} (H_n^0 \Sigma_n^0 H_n^0)^{-1/2} \sqrt{n} h(\lambda_n) = (HVH')^{-1/2} \delta
\]

and the result follows. \( \square \)
8. THE MAXIMUM LIKELIHOOD ESTIMATOR

Leaving out the details, we now show how Gallant and Holly (1978) results can be derived from the theory we have presented.

The basic relation is

\[ g(y_t, x_t, \theta^0) = u_t \quad t = 1, 2, \ldots, n \]  \hspace{1cm} (8.1)

where the residual sequence \( (u_t)_{t>1} \) is the realization of an independently and identically distributed stochastic process. The common probability law for this process has mean zero, variance-covariance matrix \( \Sigma^* \) positive definite, and is absolutely continuous with respect to the Lebesgue measure on \( \mathbb{R}^m \). The parameter \( \theta^0 \) is a vector in \( \Theta \).

**Definition 8.1**

Let \( \Sigma^{m \times m} = \sigma_{ij} \) be a symmetric matrix and \( \sigma = (\sigma_{11}, \sigma_{12}, \sigma_{22}, \ldots, \sigma_{1m}, \sigma_{2m}, \ldots, \sigma_{mm})' \). The notation \( \Sigma(\sigma) = \Sigma \) will be used.

Let \( \sigma^* \) satisfy \( \Sigma(\sigma^*) = \Sigma^* \). We can write

\[ (\Sigma(\sigma^*))^{-\frac{1}{2}} g(y_t, x_t, \theta^0) = (\Sigma(\sigma^*))^{-\frac{1}{2}} u_t \quad t = 1, 2, \ldots, n . \]

Let \( e_t = (\Sigma(\sigma^*))^{-\frac{1}{2}} u_t \) and \( G(y, x, \gamma^0) = (\Sigma(\sigma^*))^{-\frac{1}{2}} g(y, x, \gamma^0) \), \( \gamma^0 = (\theta^0, \sigma^*) \). Then (8.1) is equivalent to

\[ G(y_t, x_t, \gamma^0) = e_t \quad t = 1, 2, \ldots, n \]

where the \( e \)'s have mean zero and identity variance-covariance matrix.

Also \( \gamma^0 \in \Gamma \), where \( \Gamma \) is the cartesian product of \( \Theta \) with the set of all \( \sigma \)'s such that \( \Sigma(\sigma) \) is positive definite. Denote by \( p(e) \) the common density for the \( e \)-sequence. If the conditions of the
implicit function theorem are satisfied for G and Assumption 1 hold, for every pair \((x, \gamma)\) in \(X \times \Gamma\), we can define a density

\[
p(y|x, \gamma) = \left. \frac{\partial G}{\partial y} (y, x, \gamma) \right| p(G(y, x, \gamma)) .
\]

Let

\[
s(y, x, \lambda) = \ln p(y|x, \lambda) , \quad \lambda \in \Gamma .
\]

Then

\[
S_n(\lambda) = \frac{1}{n} \sum_{t=1}^{n} \ln p(y_t|x_t, \lambda) ,
\]

\[
\int \int s(f(e, x, \gamma^0), x, \lambda) \mu(de, dx) = \int \int \ln p(y|x, \lambda)p(y|x, \gamma^0) d\mu p_X(dx) . \tag{8.2}
\]

Assume that (8.2) has a unique maximum at \(\lambda = \gamma^0\). Under Assumptions 1 through 8, if \(\hat{\lambda}_n = (\hat{\theta}_n, \hat{\sigma}_n)\) maximizes \(S_n(\lambda)\), \(\hat{\theta}_n\) is strongly consistent for \(\theta^0\), \(\hat{\sigma}_n\) is strongly consistent for \(\sigma^*\), and

\[
\sqrt{n} \begin{bmatrix} \hat{\theta}_n - \theta^0 \\ \hat{\sigma}_n - \sigma^* \end{bmatrix} \xrightarrow{D} N(0, J^{-1})
\]

since \(J = I\) in the maximum likelihood context. The matrix \(J^{-1}\) is estimated by \(\hat{J}^{-1}_n\) (or \(\hat{I}^{-1}_n\)).

If \(h(\lambda)\) is continuously differentiable and its Jacobian has full rank at \(\lambda = \gamma^0\), we may test \(H_0: h(\gamma^0) = 0\) against \(H_A: h(\gamma^0) \neq 0\).

Suppose that \(\sigma^*\) is considered as a nuisance parameter for hypothesis.
testing so that the only functions \( h \) of interest are those of the form \( h(\lambda) = \tilde{h}(\theta) \). Then, with the notation of Chapter 7, \( H(\lambda) = (\tilde{H}(\theta), 0) \),
\( \tilde{H}(\theta) = \frac{\partial \tilde{h}}{\partial \theta} (\theta) \). The Wald's test statistic, for example, becomes

\[
W_n = n(\tilde{h}(\theta_n))'\left(\tilde{H}(\theta_n)\tilde{J}^{-1}_n \tilde{H}(\theta_n)'\right)^{-1}(\tilde{h}(\theta_n))
\]

where

\[
\tilde{J}^{-1}_n = \begin{bmatrix}
\tilde{j}^{-1}_{\theta \theta} & \tilde{j}^{-1}_{\theta \sigma} \\
\tilde{j}^{-1}_{\sigma \theta} & \tilde{j}^{-1}_{\sigma \sigma}
\end{bmatrix}
\]

the partition being conformable to the dimensions of \( \theta \) and \( \sigma \) (clearly the matrix \( \tilde{I}_n \) could also be used). Under the null hypothesis, for \( n \) large, \( W_n \) has a distribution that can be approximated by a chi-square with degrees of freedom determined by the dimension of the vector function \( \tilde{h} \). To approximate power, \( \gamma^0 \) must be specified. By (8.2), for the specification \( \gamma^0_n, \lambda^0_n = \theta^0_n \), so that proposition 7.4.1 may be used to approximate the noncentrality parameter \( \alpha \) of the asymptotic non-null distribution of \( W_n \) (also of \( R_n \) and \( L_n \)).

At this point we remark that we have diverged a little from Gallant and Holly (1978). The tests derived in their paper are based on the concentrated likelihood technique. We have not done this here, although possible and straightforward along their lines. In this respect it is interesting to observe that explicit knowledge of the limiting law of \( \sqrt{n} (\hat{\theta}_n - \theta^0) \) is not necessary to derive the asymptotic distribution of any of the three test statistics considered in Chapter 7.
9. THE SINGLE EQUATION EXPLICIT MODEL

Assume that responses $y_t$ to inputs $x_t$ are generated according to the structure

$$y_t = g(x_t, \theta^0) + u_t \quad t = 1, 2, \ldots, n.$$  \hspace{1cm} (9.1)

The parameter $\theta^0$ is an interior point of $\Theta$ which is a compact subset of $\mathbb{R}^l$. The function $g$ is continuous and, for each $x$ fixed, twice continuously differentiable in $\theta$. The residual sequence $(u_t)_{t \geq 1}$ is the realization of an independently and identically distributed sequence of random variables, each with mean zero and variance $(\sigma^*)^2$, $\sigma^* > 0$.

This model can be transformed to the standard form of Chapter 3 as follows. Let

$$G(y, x, \gamma^0) = \frac{y - g(x, \theta^0)}{\sigma^*}, \quad \gamma^0 = (\theta^0, \sigma^*).$$

Then (9.1) is equivalent to

$$G(y_t, x_t, \gamma^0) = e_t \quad t = 1, 2, \ldots, n$$

where $e_t = u_t/\sigma^*$. Here the reduced form is

$$y = f(e, x, \gamma^0) = \sigma^* e + g(x, \theta^0).$$

Define

$$s(y, x, \theta) = -\frac{1}{2}(y - g(x, \theta))^2.$$
Then

\[ S_n(\theta) = -\frac{1}{2n} \sum_{t=1}^{n} (y_t - g(x_t, \theta))^2. \]

Notice that the definition of \( s \) above implies that \( \sigma^* \) is being treated as a nuisance parameter, i.e., no interest is being focused on \( \sigma^* \) for estimation purposes.

Let Assumptions 3 and 5 hold. We have:

\[
\int \int s(f(e, x, y^0), x, \theta) \mu(\text{de}, dx) = \frac{(q^0)^2}{2} - \frac{1}{2} \int (g(x, \theta^0) - g(x, \theta))^2 P_X(dx).
\]

Assumption 6 will be satisfied if

\[
\int (g(x, \theta^0) - g(x, \theta))^2 P_X(dx)
\]

has a unique minimum at \( \theta = \theta^0 \). Assume that this is the case. If \( \hat{\theta}_n \) maximizes \( S_n(\theta) \) and the integrability conditions of Chapters 5 and 6 are satisfied then \( \hat{\theta}_n \) is strongly consistent for \( \theta^0 \) and \( \sqrt{n} (\hat{\theta}_n - \theta^0) \xrightarrow{L} N(0, J^{-1}J^{-1}) \). For illustration purposes we will compute \( J \) and \( I \). Clearly

\[
\frac{\partial^2 s}{\partial \theta \partial \theta} (y, x, \theta) = (y-g(x, \theta)) \frac{\partial^2 g}{\partial \theta \partial \theta} (x, \theta),
\]

\[
\frac{\partial^2 s}{\partial \theta^2} (y, x, \theta) = -\frac{\partial \theta}{\partial \theta} (x, \theta) \frac{\partial^2 g}{\partial \theta^2} (x, \theta) + (y-g(x, \theta)) \frac{\partial^2 g}{\partial \theta^2} (x, \theta).
\]
It follows that

\[
J = \int_X \frac{\partial^2 g}{\partial \theta^2} (x, \theta^0) \frac{\partial^2 g}{\partial \theta' \partial \theta'} (x, \theta^0) P_X(dx),
\]

\[
I = (\sigma^*)^2 \int_X \frac{\partial g}{\partial \theta} (x, \theta^0) \frac{\partial g}{\partial \theta'} (x, \theta^0) P_X(dx).
\]

Thus, the asymptotic variance covariance matrix of $\sqrt{n} (\hat{\theta}_n - \theta^0)$ is

\[
(\sigma^*)^2 \left( \int_X \frac{\partial g}{\partial \theta} (x, \theta^0) \frac{\partial g}{\partial \theta'} (x, \theta^0) P_X(dx) \right)^{-1}
\]

consistently estimated by $\hat{J}^{-1} \hat{I}^{-1}$ (Theorem 6.2).

Since $J$ is not equal to $I$ the likelihood ratio test cannot be performed. Of course the reason for this difference lies in the choice of the objective function. If the common distribution of the residual variables were absolutely continuous with respect to some measure on the real line, and $s$ defined via the likelihood, $I$ would be equal to $J$, as mentioned in Chapters 6 and 8. Consider the normal case, for example. Let

\[
s(y, x, \lambda) = -\log \sigma - \frac{1}{2\sigma^2} (y - g(x, \theta))^2 \quad \lambda = (\theta, \sigma)
\]

and assume that the identification condition (8.2) of Chapter 8 is satisfied. Here
\[
\frac{\partial s}{\partial \lambda} (y, x, \theta, \sigma) = \left[ \frac{1}{\sigma} (y-g(x, \theta) \frac{\partial g}{\partial \theta} (x, \theta) \right] \\
- \frac{1}{\sigma} + \frac{(y-g(x, \theta))^2}{\sigma^3} \\
\frac{\partial^2 s}{\partial \lambda' \partial \lambda} (y, x, \theta, \sigma) = \left[ -\frac{1}{\sigma^2} \frac{\partial g}{\partial \theta} \frac{\partial g}{\partial \theta} + \frac{1}{\sigma} (y-g) \frac{\partial^2 g}{\partial \theta' \partial \theta} - \frac{2}{\sigma^3} (y-g) \frac{\partial g}{\partial \theta} \right] \\
- \frac{2}{\sigma^3} \frac{\partial g}{\partial \theta'} (y-g) \frac{1}{\sigma^2} - \frac{3(y-g)^2}{\sigma^4} \\
\]

where arguments were deleted in the latter expression for purposes of simplicity. Then

\[
I = J = (\sigma^*)^{-2} \begin{bmatrix}
\int_{X} \frac{\partial g}{\partial \theta} (x, \theta^0) \frac{\partial g}{\partial \theta'} (x, \theta^0) P_X(dx) & 0 \\
0 & 2
\end{bmatrix}
\]

where \( \hat{\lambda}_n = (\hat{\theta}_n, \hat{\sigma}_n) \) maximizes \( S_n(\lambda) \). The estimators \( \hat{\theta}_n \) and \( \hat{\sigma}_n \) are asymptotically independent since \( J^{-1} \) is block diagonal. These results do not hold, for the same function \( s \), if the distribution generating the random components is not normal.

Specification of parameters for power purposes should be made, in both examples, as in Chapter 8. Notice that for the first example a test of \( H_0: h(\theta^0) = 0 \) against \( H_A: h(\theta^0) \neq 0 \) is the only one permitted.
10. THE SEEMINGLY UNRELATED NONLINEAR MODEL

The seemingly unrelated nonlinear model is defined by a set of \( m \) univariate nonlinear regressions

\[
y_{at} = g_a(x_{at}, \theta_a^0) + u_{at} \quad t = 1, 2, \ldots, n; \quad a = 1, 2, \ldots, m.
\]

We may write

\[
y_t = g(x_t, \theta^0) + u_t \quad t = 1, 2, \ldots, n
\]

(10.1)

where

\[
y_t = (y_{1t}, \ldots, y_{mt})',
\]

\[
x_t = (x_{1t}, \ldots, x_{mt})',
\]

\[
u_t = (u_{1t}, \ldots, u_{mt})',
\]

\[
\theta^0 = (\theta_1^0, \ldots, \theta_m^0)',
\]

\[
g(x_t, \theta^0) = (g_1(x_{1t}, \theta_1^0), \ldots, g_m(x_{mt}, \theta_m^0)').
\]

The parameter \( \theta^0 \) is an interior point of \( \Theta \) which is a compact subset of \( \mathbb{R}^q \). The function \( g \) is continuous and, for each \( x \) fixed, twice continuously differentiable in \( \theta \). The residual sequence \((u_t)_{t \geq 1}\) is a realization of an independently and identically distributed sequence of random variables each with mean zero and variance covariance matrix \( \Sigma^* \) positive definite.
In the context of Definition 8.1, let $\sigma^*$ and $\tau^*$ be such that

$$\Sigma(\sigma^*) = \Sigma^*$$ and $$\Sigma(\tau^*) = (\Sigma^*)^{-1}. \text{ Let}$$

$$G(y, x, \gamma^0) = (\Sigma(\sigma^*))^{-1/2}(y - g(x, \theta^0)), \quad \gamma^0 = (\theta^0, \sigma^*).$$

Then (10.1) is equivalent to

$$G(y_t, x_t, \gamma^0) = e_t \quad t = 1, 2, \ldots, n$$

where $e_t = (\Sigma(\sigma^*))^{-1/2} u_t$. The reduced form is

$$y = f(e, x, \gamma^0) = (\Sigma(\sigma^*))^{1/2} e + g(x, \theta^0).$$

Define

$$s(y, x, \tau, \theta) = -\frac{1}{2}(y - g(x, \theta))' \Sigma(\tau)(y - g(x, \theta)).$$

For the seemingly unrelated nonlinear model, an initial strongly consistent estimator $\hat{\tau}_n$ of $\tau^*$ is available (Gallant (1975c)). It is obtained as follows. For each of the $m$ regressions compute the residual vector $\hat{u}_a$ of the corresponding nonlinear least squares fit. Form the elements $\hat{\sigma}_{\alpha \beta} = \frac{1}{n} \hat{u}_\alpha' \hat{u}_\beta$ and let $\hat{\Sigma} = (\hat{\sigma}_{\alpha \beta}).$ The estimator $\hat{\tau}_n$ is such that $\Sigma(\hat{\tau}_n) = (\hat{\Sigma})^{-1}.$ This estimator also satisfies Assumption 4, i.e., $\sqrt{n}(\hat{\tau}_n - \tau^*)$ is bounded in probability. Thus, $s$ induces

$$S_n(\theta) = -\frac{1}{2n} \sum_{t=1}^{n} \frac{1}{n} \sum_{t=1}^{n} (y_t - g(x_t, \theta))' \Sigma(\hat{\tau}_n)(y_t - g(x_t, \theta)).$$
Let Assumptions 3 and 5 hold. We have

\[
\int \int_{\mathcal{E} \times \mathcal{X}} s(f(e,x,y^0),x,\tau^*,\theta) \mu(\text{de,dx})
\]

\[
= -\frac{m}{2} - \frac{1}{2} \int_{\mathcal{X}} \left( g(x,\theta) - g(x,\theta^0) \right) (\Sigma^*)^{-1} \left( g(x,\theta) - g(x,\theta^0) \right) P_{\mathcal{X}}(\text{dx}) .
\]

Assumption 6 will be satisfied if

\[
\int_{\mathcal{X}} \left( g(x,\theta) - g(x,\theta^0) \right) (\Sigma^*)^{-1} \left( g(x,\theta) - g(x,\theta^0) \right) P_{\mathcal{X}}(\text{dx})
\]

has a unique minimum at \( \theta = \theta^0 \). Assume that this is the case.

The following derivatives are useful to obtain the matrices \( I \) and \( J \) as well as to verify the two integral conditions of Assumption 8. Clearly,

\[
\frac{\partial s}{\partial \theta} (y,x,\tau,\theta) = \left[ \frac{\partial g}{\partial \theta} (x,\theta) \right]' \Sigma(\tau)(y-g(x,\theta)) .
\]

To compute the Jacobian matrix of \( \frac{\partial s}{\partial \theta} \), write

\[
\frac{\partial s}{\partial \theta} (x,\theta) = \left[ g_1(x,\theta),... g_k(x,\theta) \right]'.
\]

By definition,

\[
\frac{\partial^2 s}{\partial \theta^2} (y,x,\tau,\theta) =
\begin{bmatrix}
\frac{\partial}{\partial \theta} \left[ \frac{\partial}{\partial \theta} (x,\theta) \Sigma(\tau)(y-g(x,\theta)) \right]
\vdots
\frac{\partial}{\partial \theta} \left[ \frac{\partial}{\partial \theta} (x,\theta) \Sigma(\tau)(y-g(x,\theta)) \right]
\end{bmatrix}
\]
\[
\begin{bmatrix}
\frac{\partial g_f}{\partial \theta_f}(x, \theta)
\end{bmatrix}' \Sigma(\tau) \frac{\partial g_f}{\partial \theta_f}(x, \theta) + \\
(y-g(x, \theta)) \Sigma(\tau) \frac{\partial^2 g_f}{\partial \theta^2}(x, \theta)
\end{bmatrix}
\]

Also,

\[
\frac{\partial^2 s}{\partial \tau' \partial \theta} (y, x, \tau, \theta) = \\
\begin{bmatrix}
\frac{\partial}{\partial \tau'} (\bar{g}_{i1}'(x, \theta) \Sigma(\tau)(y-g(x, \theta))) \\
\vdots \\
\frac{\partial}{\partial x} (\bar{g}_{i1}'(x, \theta) \Sigma(\tau)(y-g(x, \theta)))
\end{bmatrix}
\]

The typical element in the \(i^{th}\) row of \(\frac{\partial^2 s}{\partial \tau' \partial \theta}\) is the sum of two elements of the form \(\bar{g}_{ik}'(x, \theta)(y-g(x, \theta))_{j}\), where \(\bar{g}_{ik}'(x, \theta)\) is the \(k^{th}\) component of \(\bar{g}_{i1}'(x, \theta)\) and \((y-g(x, \theta))_{j}\) is the \(j^{th}\) component of \(y-g(x, \theta)\).

It follows that \(\frac{\partial s}{\partial \theta}\) and \(\frac{\partial^2 s}{\partial \tau' \partial \theta}\) integrates to zero with respect to \(p_E\) and \(\nu\), respectively, and also

\[
I = J = \int_{\mathbb{R}} \left[ \frac{\partial g_f}{\partial \theta_f}(x, \theta^0) \right]' (\Sigma)^{-1} \left[ \frac{\partial g_f}{\partial \theta_f}(x, \theta^0) \right] P_X(dx).
\]

If \(\hat{\theta}_n\) maximizes \(S_n(\lambda)\) and the integrability conditions of Chapters 5 and 6 are satisfied, then \(\hat{\theta}_n\) is strongly consistent for \(\theta^0\) and \(\sqrt{n}(\hat{\theta}_n-\theta^0) \overset{L}{\longrightarrow} N(0, J^{-1}).\) Either \(\hat{J}_n\), or \(\hat{I}_n\) can be used to estimate the asymptotic variance covariance matrix.
Consider the problem of testing $H_0: h(\theta^0) = 0$ against

$H_A: h(\theta^0) \neq 0$, where $h$ is continuously differentiable with Jacobian

of full row rank at $\theta^0$. Since $I = J$ any of the test statistics

presented in Chapter 7 can be used. For power purposes a specification

of $\gamma$ implies a specification of $\tau$ as well as of $\lambda$ (which equals $\theta$

in this case), so that proposition 7.4.1 may be used. Notice that here

we may well use the elements $\hat{\sigma}_{a\beta} = \frac{1}{n} \hat{u}_a \hat{u}_{\beta}$ to specify the second

component of $\gamma$ so that the specification of $\tau$ is $\hat{\tau}_n$. 

11. LIST OF REFERENCES


_______. (1977), The maximum likelihood estimator and the nonlinear three stage least squares estimator in the general nonlinear simultaneous equation model. Econometrica 45, 955-968.


Wesler, O. (1979), Personal communication.