NONLINEAR STATISTICAL MODELS

by

A. Ronald Gallant

CHAPTER 3. A Unified Asymptotic Theory of Nonlinear Statistical Models

This preprint is circulated for discussion purposes. Comments and notification of any errors sent to the address below will be appreciated.

A. Ronald Gallant
Institute of Statistics
North Carolina State University
P. O. Box 5457
Raleigh, NC 27650 USA

Phone: 919-737-2531

Additional copies may be ordered from the Institute of Statistics at a price of $15.00for USA delivery; postage will be charged for overseas orders.

NONLINEAR STATISTICAL MODELS

Table of Contents

Chapter 3. A Unified Asymptotic Theory of Nonlinear Statistical Models

After reading a few articles in the nonlinear models literature one begins to notice that each discussion follows roughly the same lines as the classical treatment of maximum likelihood estimation. There are some technical problems having to do with simultaneously conditioning on the independent variables and subjecting the true parameter to a Pitman drift which prevent the use of the classical methods of proof, but the basic impression of similarity is correct. An estimator -- be it nonlinear least squares, three-stage nonlinear least squares, or whatever -- is the solution of an optimization problem. And the objective function of the optimization problem can be treated as if it were the log-likelihood to derive the Wald test statistic, the likelihood ratio test statistic, and Rao's efficient score statistic. Their asymptotic null and non-null distributions can be found using arguments fairly similar to the classical maximum likelihood arguments. In this chapter we exploit these observations and develop a unified asymptotic theory for nonlinear models. That which escapes this unification is that which has objective function which is not twice continually differentiable with respect to the parameters - minimum absolute deviations regression for example.

The model that generates the data need not be the same as the model that was presumed to define the optimization problem. Thus, these results can be used to obtain the asymptotic behavior of inference procedures under specification error. For example, it is not necessary to resort to Monte Carlo simulation to determine if an exponential fit is robust against other plausible growth models. The asymptotic approximations we give here will provide an analytic answer to the question, sufficiently accurate for most purposes.

An early version of this chapter appeared as Burguete, Gallant, and Souza (1982) together with comment by Huber (1982), Phillips (1982), and White (1982). This chapter differs from the earlier work in that the Pitman drift assumption is isolated from the results on estimation. See especially Phillips (1982) Comment and the Reply as to the subtle differences this can make.

1. INTRODUCTION

An estimator is the solution of an optimization problem. It is necessary to divide these optimization problems into two groups and study these groups separately. Afterwards, one can ignore this classification and study inference in unified fashion. These two groups are least mean distance estimators and method of moments estimators. We shall define these in turn.

Multivariate nonlinear least squares is an example of a least mean distance estimator. The estimator for the model

$$y_t = f(x_t,\theta) + e_t \qquad t = 1, 2, \ldots, n$$

where y_t is an M-vector is computed as follows. Firstly, least squares residuals $\hat{e}_{it}$ are obtained by fitting the univariate models

$$y_{it} = f_i(x_t,\theta) + e_{it} \qquad i = 1, 2, \ldots, M\,;\; t = 1, 2, \ldots, n$$

individually by least squares. Let $\hat{e}_t = (\hat{e}_{1t}, \hat{e}_{2t}, \ldots, \hat{e}_{Mt})'$ and

$$\hat{\tau} = (1/n)\,\Sigma_{t=1}^{n}\,\hat{e}_t\,\hat{e}_t'\;.$$

The multivariate nonlinear least squares estimator is that value $\hat{\theta}$ which minimizes

$$(1/n)\Sigma_{t=1}^{n}\,\tfrac{1}{2}[y_t - f(x_t,\theta)]'(\hat{\tau})^{-1}[y_t - f(x_t,\theta)]\;.$$

A general description of estimators of this type is: a least mean distance estimator is that value $\hat{\lambda}_n$ which minimizes an objective function of the form

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

The literature subsumed by this definition is: Single equation nonlinear least squares - Jennrich (1969), Malinvaud (1970a), Gallant (1973, 1975a, 1975b). Multivariate nonlinear least squares - Malinvaud (1970b), Gallant (1975c), Holly (1978). Single equation and multivariate maximum likelihood - Malinvaud (1970b), Barnett (1976), Holly (1978). Maximum likelihood for simultaneous systems - Amemiya (1977), Gallant and Holly (1980). M-estimators - Balet-Lawrence (1975), Grossman (1976), Ruskin (1978). Iteratively rescaled M-estimators - Souza and Gallant (1979).

Two-stage nonlinear least squares is an example of a method of moments estimator. The estimator for the αth equation

$$q_\alpha(y_t,x_t,\theta) = e_{\alpha t} \qquad t = 1, 2, \ldots, n$$

of a simultaneous system of M such equations --- y_t is an M-vector --- is computed as follows. One chooses instrumental variables z_t as functions of the exogenous variables x_t. Theoretical discussions of this choice consume much of the literature, but the most frequent choice in applications is low order monomials in x_t, viz.

$$z_t = (x_1,x_1^2,x_2,x_2^2,x_1\ x_2,x_3,\ldots)_t' \ .$$

The moment equations are

$$m_n(\theta) = (1/n)\Sigma_{t=1}^{n} z_t q_\alpha(y_t,x_t,\theta)$$

and the true value θ^* of θ is presumed to satisfy $\mathcal{E}m_n(\theta^*) = 0$. (Note that $q_\alpha(y_t,x_t,\theta)$ is a scalor and z_t is a vector.) The two-stage least squares estima is defined as the value $\hat{\theta}$ which minimizes

$$s_n(\theta) = \tfrac{1}{2}m_n'(\theta)[(1/n)\Sigma_{t=1}^{n} z_t, z_t']^{-1} m_n(\theta) \ .$$

A general description of estimators of this type is as follows. Define moment equations

$$m_n(\lambda) = (1/n)\Sigma_{t=1}^{n} m(y_t, x_t, \hat{\tau}_n, \lambda)$$

and a notion of distance

$$d(m, \hat{\tau}_n)$$

where we permit a dependence on a random variable $\hat{\tau}_n$ via the argument τ in $m(y,x,\tau,\lambda)$ and $d(m,\tau)$ so as to allow preliminary estimates of nuisance parameters as in three-stage least squares. The estimator is that $\hat{\lambda}_n$ which minimizes

$$s_n(\lambda) = d[m_n(\lambda), \hat{\tau}_n] \ .$$

Estimators which are properly thought of as method of moment estimators, in the sense that they can be posed no other way, are: The Hartley-Booker estimator - Hartley and Booker (1965). Scale invariate M-estimators - Ruskin (1978). Two-stage nonlinear least-squares estimators - Amemiya (1974). Three-stage nonlinear least-squares estimators - Jorgenson and Laffont (1974), Amemiya (1977), Gallant and Jorgenson (1979).

In both least mean distance estimation and method of moments estimation, one is led to regard an estimator as the value $\hat{\lambda}_n$ which minimizes an objective function $s_n(\lambda)$. This objective function depends on the sample $\{(y_t, x_t) : t = 1, 2, \ldots, n\}$ and possibly on a preliminary estimator $\hat{\tau}_n$ of some nuisance parameters. Now the negative of $s_n(\lambda)$ may be treated as if it were a likelihood function and the Wald test statistic W_n, the likelihood ratio test statistic

L_n, and Rao's efficient score test statistic R_n may be derived for a null hypothesis H: $h(\lambda) = 0$ against its alternative A: $h(\lambda) \neq 0$. Almost all of the inference procedures used in the analysis of nonlinear statistical models can be derived in this way. It is only a matter of finding the appropriate objective function $s_n(\lambda)$.

We emerge from this discussion with an interest in four statistics --- $\hat{\lambda}_n$, W_n, L_n, R_n --- all of which depend on $s_n(\lambda)$. We should like to find their asymptotic distribution in three cases: the null case where the model is correctly specified and the null hypothesis $h(\lambda) = 0$ holds, the non-null case where the model is correctly specified and the null hypothesis is violated, and in the case where the model is misspecified. By misspecification, one has in mind the following. The definition of an objective function $s_n(\lambda)$ which defines the four statistics of interest is motivated by a model and assumptions on the error distribution. For example, the multivariate nonlinear least-squares estimator is predicated on the assumption that the data follow the model

$$y_t = f(x_t,\theta) + e_t\,, \qquad t = 1, 2, \ldots, n$$

and that the errors have mean zero. Misspecification means that either the model assumption or the error assumption or both are violated. We find that we can obtain an asymptotic theory for all three cases at once by presuming that the data actually follow the multivariate implicit model

$$q(y_t,x_t,\gamma_n^o) = e_t \qquad t = 1, 2, \ldots, n$$

where y, q, and e are M-vectors and the parameter γ may be infinite dimensional. That is, we obtain our results with misspecification and violation of the null hypothesis presumed throughout and then specialize to consider correctly specified null and non-null situations. The following results are obtained.

The least mean distance estimator $\hat{\lambda}_n$, the estimator which minimizes

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

is shown to be asymptotically normally distributed with a limiting variance-covariance matrix of the form $\mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}$. Consistent estimators $\hat{\mathcal{I}}_n$ and $\hat{\mathcal{J}}_n$ are set forth. Two examples --- an M-estimator and an iteratively rescaled M-estimator --- are carried throughout the development to illustrate the regularity conditions and results as they are introduced.

Next, method of moments estimation is taken up. The method of moments estimator $\hat{\lambda}_n$, the estimator that minimizes

$$s_n(\lambda) = d[m_n(\lambda), \hat{\tau}_n] ,$$

is shown to be asymptotically normally distributed with a limiting variance-covariance matrix of the form $\mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}$. Again, consistent estimators $\hat{\mathcal{I}}_n$ and $\hat{\mathcal{J}}_n$ are set forth. The example carried throughout the discussion is a scale invariant M-estimator.

Both analyses --- least mean distance estimation and method of moments estimation --- terminate with the same conclusion: $\hat{\lambda}_n$ minimizing $s_n(\lambda)$ is asymptotically normally distributed with a limiting variance-covariance matrix that may be estimated consistently by using $\hat{\mathcal{J}}_n$ and $\hat{\mathcal{I}}_n$ as intermediate statistics. As a result, an asymptotic theory for the test statistics W_n, L_n, and R_n can be developed in a single section, Section 5, without regard to whether the source of the objective function $s_n(\lambda)$ was least mean distance estimation or method of moments estimation. The discussion is illustrated with a misspecified nonlinear regression model fitted by least squares.

Observe that a least mean distance estimator may be cast into the form of a method of moments estimator by putting

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

because $\hat{\lambda}$ which minimizes

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

solves

$$(1/n)\Sigma_{t=1}^{n}(\partial/\partial\lambda)s(y_t,x_t,\hat{\tau}_n,\lambda) = 0 .$$

If one's only interest is the asymptotic distribution of $\hat{\lambda}_n$, then posing the problem as a method of moments estimator is the more convenient approach as algebraic simplifications of the equations $m_n(\lambda) = 0$ prior to analysis can materially simplify the computation of the parameters of the asymptotic distribution. However, one pays two penalties for this convenience: the problem is no longer posed in a way that permits the use of the statistic L_n , and consistency results are weaker.

2. THE DATA GENERATING MODEL AND LIMITS OF CESARO SUMS

The objective is to find asymptotic approximations in situations such as the following. An analysis is predicted on the assumption that the data were generated according to the model

$$y_t = f(x_t,\lambda) + e_t \qquad t = 1,2,\ldots, n$$

when actually they were generated according to

$$y_t = g(x_t) + e_t \qquad t = 1,2,\ldots, n \ .$$

One estimates λ by $\hat{\lambda}_n$ that minimizes $s_n(\lambda)$ over the estimation space Λ and tests H: $\lambda = \lambda^*$ by, say,

$$W_n = n(\hat{\lambda}_n - \lambda^*)'(\hat{\mathcal{J}}^{-1}\,\hat{\mathcal{I}}\,\hat{\mathcal{J}}^{-1})^{-1}(\hat{\lambda}_n - \lambda^*) \ .$$

The estimator $\hat{\lambda}_n$ is estimating a value λ° induced by f(x) which is computed according to formulas given later. Thus, one is actually testing the null hypothesis H: $\lambda^\circ = \lambda^*$. Depending on the context, a test of H: $\lambda^\circ = \lambda^*$ when the data is generated according to

$$y_t = g(x_t) + e_t \qquad t = 1,2,\ldots, n$$

and not according to

$$y_t = f(x_t,\lambda) + e_t \qquad t = 1,2,\ldots, n$$

may or may not make sense. In order to make a judgement as to whether the inference procedure is sensible it is necessary to have the (asymptotic approximation to the) sampling distribution of W_n.

A problem in deriving asymptotic approximations to the sampling distribution of W_n is that if $\lambda^\circ \neq \lambda^*$ then W_n will reject the null hypothesis with probability one as n tends to infinity whence its limiting distribution is degenerate. The classical solution to this problem is to index the parameter as λ_n° and subject it to a rate of drift such that $\sqrt{n}(\lambda_n^\circ - \lambda^*)$ converges to a finite limit, called a Pitman drift. Thus, we need some mechanism for subjecting the true model g(x) to drift so as to induce the requisite drift on λ_n°.

One possible drift mechanism is the following. Suppose that the independent variables are confined to a compact set $\mathcal{X}$ and that $f(x,\lambda^*)$ is continuous on $\mathcal{X}$. Then $f(x,\lambda^*)$ has a polynomial expansion

$$f(x,\lambda^*) = \Sigma_{i=0}^{\infty} \gamma_i^* x^i$$

by the Stone-Weierstrass theorem. If the data is generated according to the sequence of models

$$g_1(x) = \gamma_0^* + \gamma_1^* x \qquad n = 1$$

$$g_2(x) = \gamma_0^* + \gamma_1^* x + \gamma_2^* x^2 \qquad n = 2$$

$$g_3(x) = \gamma_0^* + \gamma_1^* x + \gamma_2^* x^2 + \gamma_3^* x^3 \qquad n = 3$$

$$\vdots$$

then λ_n° will converge to that λ^* specified by H: $\lambda = \lambda^*$. Convergence can be accelerated so that $\lim \sqrt{n}(\lambda_n^\circ - \lambda^*)$ is finite by changing a few details (Problem 2). The natural representation of this scheme is to put

$$g_n(x) = g(x,\gamma_n^\circ) = \Sigma_{i=0}^{\infty} \gamma_{in}^\circ x^i$$

and let

$$\gamma_1^\circ = (\gamma_0^*, \gamma_1^*, 0, \ldots)$$

$$\gamma_2^\circ = (\gamma_0^*, \gamma_1^*, \gamma_2^*, 0, \ldots)$$

$$\gamma_3^\circ = (\gamma_0^*, \gamma_1^*, \gamma_2^*, \gamma_3^*, 0, \ldots)$$

$$\vdots$$

We see from this discussion that the theory should at least be general enough to accommodate data generating models with an infinite dimensional parameter space. Rather than working directly with an infinite dimensional parameter space, it is easier to let the parameter space be an abstract metric space (Γ, ρ). To specialize to the infinite dimensional case, let Γ be the collection of infinite dimensional vectors and put $\rho(\gamma, \gamma^\circ) = \Sigma_{i=0}^{\infty}|\gamma_i - \gamma_i^\circ|$ or some other convenient metric (Problem 2).
To specialize further to the finite dimensional case, let $\Gamma = R^s$ and put $\rho(\gamma,\gamma^\circ) = (\Sigma_{i=1}^{s}|\gamma_i - \gamma_i^\circ|^2)^{\frac{1}{2}}$.

Moving on to the formal assumptions, we assume that the observed data

$$(y_1,x_1)\,,\ (y_2,x_2)\,,\ (y_3,x_3)\,,\ \ldots$$

is generated according to the model

$$q(y_t,x_t,\gamma_n^\circ) = e_t \qquad t = 1, 2, \ldots, n$$

with $x_t \in \mathcal{X}$, $y_t \in \mathcal{Y}$, $e_t \in \mathcal{E}$, and $\gamma_n^\circ \in \Gamma$. The dimensions are: x_t is a k-vector, y_t and e_t are M-vectors, and (Γ, ρ) is an abstract metric space with γ_n° some point in Γ. The observed values of y_t are actually doubly indexed and form a triangular array

$$\begin{array}{lll} y_{11} & & n = 1 \\ y_{12}\ y_{22} & & n = 2 \\ y_{13}\ y_{23}\ y_{33} & & n = 3 \\ \quad\vdots & & \end{array}$$

due to the dependence of γ_n° on the sample size n. This second index will simply be understood throughout.

ASSUMPTION 1. The errors are independently and identically distributed with common distribution $P(e)$.

Obviously, for the model to make sense, some measure of central tendency of $P(e)$ ought to be zero but no formal use is made of such an assumption. If $P(e)$ is indexed by parameters, they cannot drift with sample size as may γ_n° .

The assumption appears to rule out heteroscedastic errors. Actually it does not if one is willing to presume that the error variance-covariance matrix depends on the independent variable x_t ,

$$\mathcal{C}(e_t, e_t') = \Sigma(x_t) \ .$$

Factor $\Sigma^{-1}(x_t)$ as $R'(x_t)\ R(x_t)$ and write

$$R(x_t)q(y_t, x_t, \gamma_n^\circ) = R(x_t)e_t \ .$$

Then $R(x_t)e_t$ is homoscedastic. If one is willing to assume a common distribution for $R(x_t)e_t$ as well then Assumption 1 is satisfied. Note that the actual construction of $R(x_t)$ is not required in applications as estimation is based only on the known function $s_n(\lambda)$. Similarly, many other apparent departures from Assumption 1 can be accommodated by presuming the existence of a transformation $\psi[q(y,x,\gamma_{(1)}),x,\gamma_{(2)}]$ that will yield residuals that satisfy Assumption 1 .

The model is supposed to describe the behavior of some 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 $\mathcal{E}$ onto $\mathcal{Y}$ denoted as $Y(e,x,\gamma)$. Moreover, $Y(e,x,\gamma)$ is continuous on $\mathcal{E} \times \mathcal{X} \times \Gamma$ and Γ is compact.

It should be emphasized that it is not necessary to have a closed form expression for $Y(e,x,\gamma)$, or even to be able to compute it using numerical methods in order to use the statistical methods set forth here. Inference is based only on the known function $s_n(\lambda)$. The existence of $Y(e,x,\gamma)$ is needed but its construction is not required. This point is largely irrelevant to standard regression models but it is essential to nonlinear simultaneous equation models where $Y(e,x,\gamma)$ is often difficult to compute. Since Γ may be taken as $\{\gamma^*,\gamma_1,\gamma_2,\ldots\}$ if desired, no generality is lost by assuming that Γ is compact.

Repeatedly in the analysis of nonlinear models a Cesaro sum such as

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

must converge uniformly in (γ°,λ) to obtain a desired result. If results are to be useful in applications, the conditions imposed to insure this uniform convergence should be plausible and easily recognized as obtaining or not obtaining in an application. The conditions imposed here have evolved in Jennrich (1969), Malinvaud (1970a), Gallant (1977a), Gallant and Holly (1980), and Burguete, Gallant and Souza (1982).

As motivation for these conditions, consider the sequence of independent variables resulting from a treatment-control experiment where the response depends on the age of the experimental material. Suppose subjects are randomly selected from a population whose age distribution is $F_A(\cdot)$ and then subjected to either the treatment or the control. The observed sequence of independent variables is

$x_1 = (1, a_1)$	treatment
$x_2 = (0, a_2)$	control
$x_3 = (1, a_3)$	treatment
$x_4 = (0, a_4)$	control
$\vdots$	

Let $F_p(\cdot)$ denote the point binomial distribution with $p = \frac{1}{2}$ and set

$$d\mu(x) = dF_p(x_1) \times dF_A(x_2)$$

Then for any continuous function $f(x)$ whose expectation exists,

$$\lim_{n\to\infty}(1/n)\, \Sigma_{t=1}^{n} f(x_t) = \Sigma_{i=0}^{1} \tfrac{1}{2}\int f(i,a)\,dF_A(a) = \int_{\mathcal{X}} f(x)\, d\mu(x)$$

for almost every realization of $\{x_t\}$ by the Strong Law of Large Numbers. The null set depends on the function $f(x)$ which would be an annoyance as the discussion flows more naturally if one has the freedom to hold a realization of $\{x_t\}$ fixed while permitting $f(x)$ to vary over a possibly uncountable collection of functions. Fortunately, the collection of functions considered later is dominated and we can take advantage of that fact now to eliminate this dependence of the null set on $f(x)$. Consider the following consequence of the generalized Glivenko-Cantelli Theorem.

PROPOSITION 1. (Gallant and Holly, 1980) Let V_t , $t = 1, 2, \ldots$ be a sequence of independent and identically distributed s-dimensional random variables defined on a complete probability space $(\Omega, \mathcal{C}_0, P_0)$ 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 $P_0(E) = 0$ such that if $\omega \notin E$

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

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

The conclusion of this proposition describes the behavior that is required of a sequence $v_t = x_t$ or $v_t = (e_t, x_t)$. As terminology for it, such a sequence is called a Cesaro Sum Generator.

DEFINITION. (Cesaro Sum Generator; Gallant and Holly, 1980) A sequence $\{v_t\}$ of points from a Borel set $\mathcal{V}$ is said to be a Cesaro Sum Generator with respect to a probability measure ν defined on the Borel subsets of $\mathcal{V}$ and a dominating function $b(v)$ with $\int b d\nu < \infty$ if

$$\ell\text{im}_{n\to\infty}(1/n)\Sigma_{t=1}^{n} f(v_t) = \int f(v)\, d\nu(v)$$

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

We have seen that independent variables generated according to an experimental design or by random sampling satisfy this definition. Many other situations such as stratified or cluster sampling will satisfy the definition as well. We shall assume, below, that the sequence $\{x_t\}$ upon which the results are conditioned is a Cesaro Sum Generator as is almost every joint realization $\{(e_t,x_t)\}$. Then we derive the Uniform Strong Law of Large Numbers.

ASSUMPTION 3. (Gallant and Holly, 1980) Almost every realization of $\{v_t\}$ with $v_t = (e_t,x_t)$ is a Cesaro Sum Generator with respect to the product measure

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

and dominating function $b(e,x)$. The sequence $\{x_t\}$ is a Cesaro Sum Generator 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 $\int_{\mathcal{E}} \sup_{N_x} b(e,x)\, dP(e) < \infty$.

THEOREM 1. (Uniform Strong Law of Large Numbers) Let Assumptions 1 through 3 hold. Let $\langle B,\sigma\rangle$ and $\langle\Gamma,\rho\rangle$ be compact metric spaces and let $f(y,x,\beta)$ be continuous on $\mathcal{Y} \times \mathcal{X} \times B$. Let

$$|f(y,x,\beta)| \le |\, b[q(y,x,\gamma),x]| \text{ or equivalently } |f[Y(e,x,\gamma),x,\beta]| \le b(e,x)$$

for all $(y,x) \in \mathcal{Y} \times \mathcal{X}$ and all $(\beta,\gamma) \in B \times \Gamma$ where $b(e,x)$ is given by Assumption 3. Then both

$$(1/n)\ \Sigma_{t=1}^{n}\ f(y_t, x_t, \beta) \text{ and}$$

$$(1/n)\ \Sigma_{t=1}^{n} \int_{\mathcal{E}} f[Y(e, x_t, \gamma),\ x_t, \beta]\ dP(e)$$

converge uniformly to

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

over $B \times \Gamma$ except on the event E with $P_0(E) = 0$ given by Assumption 3. Recall that the uniform limit of continuous functions is continuous.

PROOF. (Jennrich, 1969) Let $v = (e,x)$ denote a typical element of $\mathcal{V} = \mathcal{E} \times \mathcal{X}$, let $\alpha = (\beta,\gamma)$ denote a typical element of $A = B \times \Lambda$, and let $\{v_t\}$ be a Cesaro Sum Generator. The idea of the proof is to use the Dominated Convergence Theorem and Cesaro summability to show that

$$h_n(\alpha) = (1/n)\ \Sigma_{t=1}^{n}\ h(v_t, \alpha)$$

where

$$h(v,\alpha) = f[Y(e,x,\gamma),\ x, \beta]$$

is an equicontinuous sequence on A. An equicontinuous sequence that has a pointwise limit on a compact set converges uniformly; see, for example, Chapter 9 of Royden (1963).

First, in order to apply Cesaro summability, we show that $\sup_{\alpha \in O} h(v,\alpha)$ and $\inf_{\alpha \in O} h(v,\alpha)$ are continuous for any $O \subset A$; they are obviously dominated by $b(e,x)$. Put $\tau(\alpha,\alpha^\circ) = [\sigma^2(\beta,\beta^\circ) + \rho^2(\gamma,\gamma^\circ)]^{\frac{1}{2}}$ whence $<A,\tau>$ is a compact metric space. Let v° in $\mathcal{V}$ and $\epsilon > 0$ be given. Let $\bar{V}$ be a compact neighborhood of v° and let $\bar{O}$ be the closure of O in $<A,\tau>$ whence $<\bar{O},\tau>$ is compact. By assumption, $h(v,\alpha)$ is continuous on $\mathcal{V} \times A$ so it is uniformly continuous on $\bar{V} \times \bar{O}$. Then there is a $\delta > 0$ such that for all $|v - v^\circ| < \delta$ and $\alpha \in \bar{O}$

$$h(v^\circ,\alpha) - \epsilon < h(v,\alpha) < h(v^\circ,\alpha) + \epsilon$$

This establishes continuity (Problem 4).

A sequence is equicontinuous if for each $\epsilon > 0$ and α° in A there is a $\delta > 0$ such that $\tau(\alpha,\alpha^\circ) < \delta$ implies $\sup_n |h_n(\alpha) - h_n(\alpha^\circ)| < \epsilon$. When each $h_n(\alpha)$ is continuous over A it suffices to show that $\sup_{n>N}|h_n(\alpha) - h_n(\alpha^\circ)| < \epsilon$ for some finite N . Let $\epsilon > 0$ and $\delta > 0$ be given and let $O_\delta = \{\alpha : \tau(\alpha,\alpha^\circ) < \delta\}$. By the Dominated Convergence Theorem and continuity

$$\begin{aligned} &\lim_{\delta \to 0} \int_{\mathcal{V}} \sup_{O_\delta} h(v,\alpha) - h(v,\alpha^\circ)\ d\nu(v) \\ &\quad = \int_{\mathcal{V}} \lim_{\delta \to 0} \sup_{O_\delta} h(v,\alpha) - h(v,\alpha^\circ)\ d\nu(v) \\ &\quad = 0 \end{aligned}$$

Then there is a $\delta > 0$ such that $\tau(\alpha,\alpha^\circ) < \delta$ implies

$$\int_{\mathcal{V}} \sup_{O_\delta} h(v,\alpha) - h(v,\alpha^\circ)\ d\nu(v) < \epsilon/2 \quad .$$

By Cesaro summability, there is an N such that $n > N$ implies

$$\sup_{O_\delta} h_n(\alpha) - h_n(\alpha^\circ) - \int_{\mathcal{V}} \sup_{O_\delta} h(v,\alpha) - h(v,\alpha^\circ)\ d\nu(v) < \epsilon/2$$

whence

$$h_n(\alpha) - h_n(\alpha^\circ) \leq \sup_{O_\delta} h_n(\alpha) - h_n(\alpha^\circ) < \epsilon$$

for all $n > N$ and all $\tau(\alpha,\alpha^\circ) < \delta$. A similar argument applied to $\inf_{O_\delta} h_n(\alpha)$ yields

$$-\epsilon < h_n(\alpha) - h_n(\alpha^\circ) < \epsilon$$

for all $n > N$ and all $\tau(\alpha,\alpha^\circ) < \delta$. This establishes equicontinuity.

To show that

$$\bar{h}_n(\alpha) = (1/n)\ \Sigma_{t=1}^{n}\ \bar{h}(x_t,\alpha)$$

where

$$\bar{h}(x,\alpha) = \int_{\mathcal{E}} f[Y(e,x,\gamma),x,\beta]\ dP(e)$$

is an equicontinuous sequence, the same argument can be applied. It is only necessary to show that $\bar{h}(x,\alpha)$ is continuous on $\mathcal{X} \times A$ and dominated by $b(x)$. Now

$$|\bar{h}(x,\alpha)| \le \int_{\mathcal{E}} |h(v,\alpha)| \; dP(e) \le \int_{\mathcal{E}} b(e,x) \; dP(e) = b(x)$$

which establishes domination. By continuity on $\mathcal{V} \times A$ and the Dominated Convergence Theorem with $\sup_{N_{x^\circ}} b(e,x)$ of Assumption 3 as the dominating function,

$$\begin{aligned} \lim_{(x,\alpha)\to(x^\circ,\alpha^\circ)} \bar{h}(x,\alpha) &= \int_{\mathcal{E}} \lim_{(x,\alpha)\to(x^\circ,\alpha^\circ)} h(e,x,\alpha) \; dP(e) \\ &= \int_{\mathcal{E}} h(e,x^\circ,\alpha^\circ) \; dP(e) \\ &= \bar{h}(x^\circ,\alpha^\circ) \; . \end{aligned}$$

This establishes continuity. ▯

In typical applications; an error density $p(e)$ and a Jacobian

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

are available. With these in hand, the conditional density

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

may be used for computing limits of Cesaro sums since

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

The choice of integration formulas is dictated by convenience.

The main use of the Uniform Strong Law is in the following type of argument:

$$\ell im_{n\to\infty}\ \hat{\lambda}_n = \lambda^* ,$$

$$\ell im_{n\to\infty}\ \sup_\Lambda |s_n(\lambda) - s^*(\lambda)| = 0 ,$$

implies $s^*(\lambda)$ continuous

$$\ell im_{n\to\infty}\ s_n(\hat{\lambda}_n) = s^*(\lambda^*)$$

because

$$|s_n(\hat{\lambda}_n) - s^*(\lambda^*)| = |s_n(\hat{\lambda}_n) - s^*(\hat{\lambda}_n) + s^*(\hat{\lambda}_n) - s^*(\lambda^*_n)|$$

$$\leq \sup_\Lambda |s_n(\lambda) - s^*(\lambda)| + |s^*(\hat{\lambda}_n) - s^*(\lambda^*_n)| .$$

We could get by with a weaker result that merely stated:

$$\ell im_{n\to\infty}\ s_n(\hat{\lambda}_n) = s^*(\lambda^*)$$

for any sequence with

$$\ell im_{n\to\infty}\ \hat{\lambda}_n = \lambda^* .$$

For the Central Limit Theorem, we shall make do with this weaker notion of convergence:

THEOREM 2. (Central Limit Theorem) Let Assumptions 1 through 3 hold. Let $<\Gamma,\rho>$ be a compact metric space; let T be a closed ball in a Euclidean space centered at τ^* with finite, nonzero radius; and let Λ be a compact subset of a Euclidean space. Let $\{\gamma^o_n\}$ be a sequence from Γ that converges to γ^* ; let $\{\hat{\tau}_n\}$ be a sequence of random variables with range in T that converges almost surely to τ^* ; let $\{\tau^o_n\}$ be a sequence from T with $\sqrt{n}(\hat{\tau}_n - \tau^o_n)$ bounded in probability; let $\{\lambda^o_n\}$ be a sequence from Λ that converges to λ^* . Let $f(y,x,\tau,\lambda)$ be a p-vector valued function such that each element of

$f(y,x,\tau,\lambda)$, $f(y,x,\tau,\lambda)\, f'(y,x,\tau,\lambda)$, and $(\partial/\partial\tau')f(y,x,\tau,\lambda)$ is continuous on $\mathcal{Y} \times \mathcal{X} \times T \times \Lambda$ and dominated by $b[q(y,x,\gamma),x]$ for all $(y,x) \in \mathcal{Y} \times \mathcal{X}$ and all $(\gamma,\tau,\lambda) \in \Gamma \times T \times \Lambda$; $b(e,x)$ is given by Assumption 3. If

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

then

$$(1/\sqrt{n})\Sigma_{t=1}^{n}[f(y_t,x_t,\hat{\tau}_n,\lambda_n^\circ) - \mu(x_t,\gamma_n^\circ,\tau_n^\circ,\lambda_n^\circ)] \xrightarrow{\mathcal{L}} N(0,I^*)$$

where

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

$$I^* = \int_{\mathcal{X}}\int_{\mathcal{E}} f[Y(e,x,\gamma^*),x,\tau^*,\lambda^*]\, f'[Y(e,x,\gamma^*),x,\tau^*,\lambda^*]\, dP(e)\, d\mu(x) - \mathcal{U}^*$$

$$\mathcal{U}^* = \int_{\mathcal{X}} \mu(x,\gamma^*,\tau^*,\lambda^*)\, \mu'(x,\gamma^*,\tau^*,\lambda^*)\, d\mu(x) \cdot$$

I^* may be singular.

PROOF. Let

$$Z(e,x,\gamma,\tau,\lambda) = f[Y(e,x,\gamma),x,\tau,\lambda] - \int_{\mathcal{E}} f[Y(e,x,\gamma),x,\tau,\lambda]\, dP(e) .$$

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

$$Z_{tn} = \ell' Z(e_t,x_t,\gamma_n^\circ,\tau_n^\circ,\lambda_n^\circ) \qquad t = 1,2,\ldots,n;\ n = 1,2,\ldots$$

Each Z_{tn} has mean zero and variance

$$\sigma_{tn}^2 = \ell' \int_{\mathcal{E}} Z(e,x_t,\gamma_n^\circ,\tau_n^\circ,\lambda_n^\circ)\, Z'(e,x_t,\gamma_n^\circ,\tau_n^\circ,\lambda_n^\circ)\, dP(e)\, \ell .$$

Putting $V_n = \Sigma_{t=1}^{n}\sigma_{tn}^2$, by Theorem 1 and the assumption that $\ell im_{n\to\infty}(\gamma_n^\circ,\tau_n^\circ,\lambda_n^\circ) = (\gamma^*,\tau^*,\lambda^*)$ it follows that $\ell im_{n\to\infty}(1/n)V_n = \ell' I^* \ell$ (Problem 5). Now $(1/n)V_n$ is the variance of $(1/\sqrt{n})\Sigma_{t=1}^{n} Z_{tn}$ and if $\ell' I^* \ell = 0$ then $(1/\sqrt{n})\Sigma_{t=1}^{n} Z_{tn}$ converges in distribution to $N(0,\ell' I^* \ell)$ by Chebyshev's inequality. Suppose, then, that $\ell' I^* \ell > 0$. If it is shown that for every $\epsilon > 0$ $\ell im_{n\to\infty} B_n = 0$ where

$$B_n = (1/n)\Sigma_{t=1}^{n} \int_{\mathcal{E}} I_{[|z|>\epsilon\sqrt{V_n}]}[\ell' Z(e,x_t,\gamma_n^o,\tau_n^o,\lambda_n^o)][\ell' Z(e,x_t,\gamma_n^o,\tau_n^o,\lambda_n^o)]^2 \, dP(e)$$

then $\ell im_{n\to\infty}(n/V_n)B_n = 0$. This is the Lindberg-Feller condition (Chung, 1974); it implies that $(1/\sqrt{n})\Sigma_{t=1}^{n} Z_{tn}$ converges in distribution to $N(0,\ell',I^*\ell)$.

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

$$\bar{B}(\gamma^*,\tau^*,\lambda^*) = \int_{\mathcal{X}}\int_{\mathcal{E}} I_{[|z|>\epsilon a]}[\ell' Z(e,x,\gamma^*,\tau^*,\lambda^*)][\ell' Z(e,x,\gamma^*,\tau^*,\lambda^*)]^2 dP(e)\, d\mu(x) .$$

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

$$I_{[|z|>\epsilon\sqrt{V_n}]}(z) \le \varphi(z) \le I_{[|z|>\epsilon a]}(z)$$

and set

$$\tilde{B}_n(\gamma,\tau,\lambda) = (1/n)\Sigma_{t=1}^{n}\int_{\mathcal{E}} \varphi[\ell' Z(e,x_t,\gamma,\tau\ \lambda)][\ell' Z(e,x_t,\gamma,\tau,\lambda)]^2 \, dP(e) .$$

By Theorem 1, $\tilde{B}_n(\gamma,\tau,\lambda)$ converges uniformly on $\Gamma \times T \times \Lambda$ to, say, $\tilde{B}(\gamma,\tau,\lambda)$. By assumption (Problem 5) $\ell im_{n\to\infty}(\gamma_n^o,\tau_n^o,\lambda_n^o) = (\gamma^*,\tau^*,\lambda^*)$ whence $\ell im_{n\to\infty}\tilde{B}_n(\gamma_n^o,\tau_n^o,\lambda_n^o) = \tilde{B}(\gamma^*,\tau^*,\lambda^*)$. Then there is an N_2 such that, for all $n > N_2$, $\tilde{B}_n(\gamma_n^o,\tau_n^o,\lambda_n^o) < \tilde{B}(\gamma^*,\tau^*,\lambda^*) + \eta/2$. But for all $n > N = \max\{N_1,N_2\}$, $B_n \le \tilde{B}_n(\gamma_n^o,\tau_n^o,\lambda_n^o)$ whence

$$B_n \le \tilde{B}_n(\gamma_n^o,\tau_n^o,\lambda_n^o) < \tilde{B}(\gamma^*,\tau^*,\lambda^*) + \eta/2 \le \bar{B}(\gamma^*,\tau^*,\lambda^*) + \eta/2 < \eta .$$

By Taylor's theorem, expanding about τ_n^o ,

$$(1/\sqrt{n})\Sigma_{t=1}^{n}\ell'[f(y_t,x_t,\hat{\tau}_n,\lambda_n^o) - \mu(x_t,\gamma_n^o,\tau_n^o,\lambda_n^o)]$$

$$= (1/\sqrt{n})\Sigma_{t=1}^{n}Z_{tn}$$

$$+ [(1/n)\Sigma_{t=1}^{n}\ell'(\partial/\partial\tau')\, f(y_t,x_t,\bar{\tau}_n,\lambda_n^o)]\,\sqrt{n}\,(\hat{\tau}_n - \tau_n^o)$$

where $\bar{\tau}_n$ lies on the line segment joining $\hat{\tau}_n$ to τ_n^o ; thus $\bar{\tau}_n$ converges almost surely to τ^* . The almost sure convergence of $(\gamma_n^o,\bar{\tau}_n,\lambda_n^o)$ to $(\gamma^*,\tau^*,\lambda^*)$ and the uniform almost sure convergence of

$$(1/n)\Sigma_{t=1}^{n}\ell'(\partial/\partial\tau)\, f(y_t,x_t,\tau,\lambda)$$

over $\Gamma \times T \times \Lambda$ given by Theorem 1 imply that $[(1/n)\Sigma_{t=1}^{n}\ell'(\partial/\partial\tau')f(y_t,x_t,\bar{\tau}_n,\lambda_n^o)]$ converges almost surely (Problem 1) to

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

Since $\sqrt{n}\,(\hat{\tau}_n - \tau_n^o)$ is bounded in probability we have that

$$(1/\sqrt{n})\Sigma_{t=1}^{n}[f(y_t,x_t,\hat{\tau}_n,\lambda_n^o) - \mu(x_t,\gamma_n^o,\tau_n^o,\lambda_n^o)]$$

$$= (1/\sqrt{n})\Sigma_{t=1}^{n}\, Z_{tn} + o_p(1)$$

$$\xrightarrow{\mathcal{L}} N(0,\ell'I^*\ell) .$$

This holds for every ℓ with $\|\ell\| = 1$ whence the desired result obtains. □

In the main, small sample regression analysis is conditional. With a model such as

$$y_t = f(x_t,\theta) + e_t \qquad t = 1,2,\ldots, n$$

the independent variables are held fixed and the sampling variation enters via the errors $e_1, e_2, \ldots ,e_n$. It seems appropriate, then, to maintain this conditioning when passing to the limit. This is what we shall do in the sequel. One fixes an infinite sequence

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

that satisfies the Cesaro summability property and all sampling variation enters via the random variables $\{e_t\}_{t=1}^\infty$. To give an unambiguous description of this conditioning, it is necessary to spell out the probability structure in detail. The reader who has no patience with details of this sort is invited to skip to the next section at this point.

We begin with an abstract probability space $(\Omega, \mathcal{A}_0, P_0)$ on which are defined random variables $\{E_t\}_{t=1}^\infty$ and $\{X_t\}_{t=1}^\infty$ which represent the errors and independent variables respectively. Nonrandom independent variables are represented in this scheme by random variables that take on a single value with probability one. A realization of the errors can be denoted by an infinite dimensional sequence

$$e_\infty = (e_1, e_2, \dots)$$

where $e_t = E_t(\omega)$ for some ω in Ω. Similarly for the independent variables

$$x_\infty = (x_1, x_2, \dots) \ .$$

Let $\mathcal{E}_\infty = \times_{t=1}^\infty \mathcal{E}$ and $\mathcal{X}_\infty = \times_{t=1}^\infty \mathcal{X}$ so that all joint realizations of the errors and independent variables take their values in $\mathcal{E}_\infty \times \mathcal{X}_\infty$ and all realizations of the independent variables take their values in $\mathcal{X}_\infty$.

Using the Daniell-Kolmogorov construction (Tucker, 1967, Section 2.3), this is enough to define a joint probability space

$$(\mathcal{E}_\infty \times \mathcal{X}_\infty \, , \ \mathcal{A}_{e,x} \, , \nu_\infty)$$

such that if a random variable is a function of (e_∞, x_∞) one can perform all computations withrespect to the more structured space $(\mathcal{E}_\infty \times \mathcal{X}_\infty, \mathcal{A}_{e,x}, \nu_\infty)$ and one is spared the trouble of tracing pre-images back to the space $(\Omega, \mathcal{A}_0, P_0)$. Similarly one can construct the marginal probability space

$$(\mathcal{X}_\infty, \mathcal{A}_x, \mu_\infty) \; .$$

Assumption 3 imposes structure on both of these probability spaces. The set on which Cesaro summability fails jointly

$$F_{e,x} = \cup_{\epsilon > 0} \cap_{j=0}^{\infty} \cup_{n=j}^{\infty} \{(e_\infty, x_\infty) : \exists \, |f(e,x)| < b(e,x) \ni |(1/n)\Sigma_{t=1}^{n} f(e_t, x_t) - \int\int f \, dP \, d\mu \, | > \epsilon\}$$

has ν_∞ measure zero. And the set on which Cesaro summability fails marginally

$$F_x = \cup_{\epsilon > 0} \cap_{j=0}^{\infty} \cup_{n=j}^{\infty} \{x_\infty : \exists \, |f(x)| < b(x) \ni |(1/n)\Sigma_{t=1}^{n} f(x_t) - \int f \, d\mu \, | > \epsilon\}$$

has μ_∞ measure zero.

By virtue of the construction of $(\mathcal{E}_\infty \times \mathcal{X}_\infty, \mathcal{A}_{e,x}, \nu_\infty)$ from countable families of random variables, there exists (Loeve, 1963, Sec. 27.2, Regularity Theorem) a regular conditional probability $P(A|x_\infty)$ connecting the joint and the marginal spaces by

$$\nu_\infty(A) = \int_{\mathcal{X}} P(A|x_\infty) \, d\mu_\infty(x_\infty) \; .$$

Recall that a regular conditional probability is a mapping of $\mathcal{A}_{e,x} \times \mathcal{X}_\infty$ into $[0,1]$ such that $P(A|x_\infty)$ is a probability measure on $(\mathcal{E}_\infty \times \mathcal{X}_\infty, \mathcal{A}_{e,x})$ for each fixed x_∞, such that $P(A|x_\infty)$ is a measurable function over $(\mathcal{X}_\infty, \mathcal{A}_x)$ for each fixed A, and such that $\int_B P(A|x_\infty) \, d\mu_\infty(x_\infty) = \nu_\infty[A \cap (\mathcal{E}_\infty \times B)]$ for every B in $\mathcal{A}_x$. The simplest example that comes to mind is to assume that $\{E_t\}_{t=1}^{\infty}$ and $\{X_t\}_{t=1}^{\infty}$ are independent families of random variables, to construct $(\mathcal{E}_\infty, \mathcal{A}_e, P_e)$, and to put

$$P(A|x_\infty) = \int_{\mathcal{E}_\infty} I_A(e_\infty, x_\infty)\, dP_e(e_\infty) \ .$$

Define the marginal conditional distribution on $(\mathcal{E}_\infty, \mathcal{G}_e)$ by

$$P_{e|x}(E|x_\infty) = P(E x \mathcal{X}_\infty | x_\infty) \ .$$

All probability statements in the sequel are with respect to $P_{e|x}(E|x_\infty)$. Assumption 1 puts additional structure on $P_{e|x}(E|x_\infty)$. It states that $P_{e|x}(E|x_\infty)$ is a product measure corresponding to a sequence of independent random variables each having common distribution P(e) defined over measurable subsets of $\mathcal{E}$. This distribution can depend on x_∞. For example, $\{e_t\}_{t=1}^\infty$ could be a sequence of independently and normally distributed random variables each with mean zero and variance-covariance matrix $\ell im_{n\to\infty}(1/n)\Sigma_{t=1}^\infty T(x_t)T'(x_t)$. But as indicated by the discussion following Assumption 1, this dependence on x_∞ is very restricted. So restricted, in fact, that we do not bother to reflect it in our notation; we do not index P of Assumption 1 by x_∞.

If all probability statements are with respect to $P_{e|x}(E|x_\infty)$ then the critical question becomes: Does the set where Cesaro summability fails conditionally at $x_\infty = x_\infty^\circ$

$$F_{e|x}^\circ = \cup_{\epsilon>0} \cap_{j=0}^\infty \cup_{n=j}^\infty \{e_\infty : \exists |f(e,x)| < b(e,x) \ni |(1/n)\Sigma_{t=1}^n f(e_t, x_t^\circ) - \iint f\, dP\, d\mu| > \epsilon\}$$

have conditional measure zero? The following computation shows that the answer is yes for almost every choice of x_∞°:

$$P_{e|x}(F^\circ_{e|x}|x^\circ_\infty) = \int_{\mathcal{E}_\infty} I_{F^\circ_{e|x}}(e_\infty)\, dP_{e|x}(e_\infty|x^\circ_\infty) \qquad \text{(marginal } |x^\circ)$$

$$= \int_{\mathcal{E}_\infty} I_{F^\circ_{e|x}\times\{x^\circ_\infty\}}(e_\infty,x^\circ_\infty)\, dP_{e|x}(e_\infty|x^\circ_\infty) \qquad \text{(marginal } |x^\circ)$$

$$= \int_{\mathcal{E}_\infty\times\mathcal{X}_\infty} I_{F^\circ_{e|x}\times\{x^\circ_\infty\}}(e_\infty,x^\circ_\infty)\, dP[(e_\infty,x_\infty)|x^\circ_\infty] \qquad \text{(joint } |x^\circ)$$

$$= \int_{\mathcal{E}_\infty\times\mathcal{X}_\infty} I_{F^\circ_{e|x}\times\{x^\circ_\infty\}}(e_\infty,x_\infty)\, dP[(e_\infty,x_\infty)|x^\circ_\infty] \qquad \text{(joint } |x^\circ)$$

$$\le \int_{\mathcal{E}_\infty\times\mathcal{X}_\infty} I_{F_{e,x}}(e_\infty,x_\infty)\, dP[(e_\infty,x_\infty)|x^\circ_\infty] \qquad \text{(joint } |x^\circ)$$

$$= P(F_{e,x}|x^\circ_\infty)\ .$$

Since

$$\nu_\infty(F_{e,x}) = \int_{\mathcal{X}} P(F_{e,x}|x^\circ_\infty)\, d\mu_\infty(x^\circ_\infty) = 0$$

we have

$$P_{e|x}(F^\circ_{e|x}|x^\circ_\infty) = 0 \ \text{ a.e. } (\mathcal{X}_\infty,\mathcal{A}_\infty,\mu_\infty)\ .$$

Since the parameter γ°_n is subject to drift, it is as well to spell out a few additional details. For each n, the conditional distribution of the dependent variables $\{y_t\}_{t=1}^n$ given x_∞ and γ°_n is defined by

$$P_n(A|x_\infty,\gamma^\circ_n) = P_{e|x}\{e_\infty \in \mathcal{E}_\infty\colon [Y(e_t,x_1,\gamma^\circ_n), \ldots ,Y(e_n,x_n,\gamma^\circ_n)] \in A|x_\infty\}$$

for each measurable subset A of $\times_{i=1}^n \mathcal{Y}$. A statement such as $\hat{\lambda}_n$ converges almost surely to λ^* means that $\hat{\lambda}_n$ is a random variable with argument $(y_1,\ldots,y_n,x_1,\ldots,x_n)$, and that $P_{e|x}(E|x_\infty) = 0$ where

$$E = \cup_{\epsilon > 0} \cap_{j=1}^{\infty} \cup_{n=j}^{\infty} \{e_{\infty} : |\hat{\lambda}_n - \lambda^*| > \epsilon\} \Big|_{(y_t, x_t) = [Y(e_t, x_t, \gamma_n^\circ), x_t]} .$$

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

$$A = (-\infty, \lambda_1] \times (-\infty, \lambda_2] \times \cdots \times (-\infty, \lambda_p]$$

it is true that

$$\ell im_{n\to\infty} P_n(\sqrt{n}(\hat{\lambda}_n - \lambda^*) \in A | x_{\infty}, \gamma_n^\circ) = \int_A dN(z|\delta, V) .$$

One may prefer an analysis that treats x_t as random rather than fixed. The theory that we shall develop is general enough to accommodate this assumption. The details are spelled out in Section 7.

We shall assume that the estimation space Λ is compact. Our defense of this assumption is that it does not cause problems in applications as a general rule and it can be circumvented on an ad hoc basis as necessary without affecting the results. We explain.

One does not wander haphazardly into nonlinear estimation. As a rule, one has need of a considerable knowledge of the situation in order to construct the model. In the computations, a fairly complete knowledge of admissible values of λ is required in order to be able to find starting values for nonlinear optimization algorithms. Thus, a statistical theory which presumes this same knowledge is not limited in its scope of applications. Most authors apparently take this position, as the assumption of a compact estimation space is more often encountered than not.

One may be reluctant to impose bounds on scale parameters and parameters that enter the model linearly. Frequently these are regarded as nuisance parameters in an application and one has little feel for what values they ought to have. Scale parameters are often computed from residuals, so start values are unnecessary, and, at least for least squares, linear parameters need no start values either (Golub and Pereya, 1973). Here, then, a compact parameter space is an annoyance.

These situations can be accommodated without disturbing in the least the results obtained here as follows. Our results are asymptotic, so if there is a compact set Λ' such that for each realization of $\{e_t\}$ there is an N where $n > N$ implies

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

then the asymptotic properties of $\hat{\lambda}_n$ are the same whether the estimation space is Λ or Λ'. For examples using this device to accommodate parameters entering linearly, see Gallant (1973). See Gallant and Holly (1980) for application to scale parameters. Other devices, such as the use of an initial consistent estimator as the start value for an algorithm which is guaranteed to converge to a local minimum of $s_n(\lambda)$, are effective as well.

PROBLEMS

1. Referring to the discussion following Theorem 2, show that if $\{X_t\}$ and $\{E_t\}$ are independent sequences of random variables, then $P_{e|x}(E|x_\infty)$ does not depend on x_∞.

2. (Construction of a Pitman drift). Consider the example of the first few paragraphs of this section where the fitted model is

$$y_t = f(x_t,\lambda) + u_t \qquad t = 1, 2, \ldots, n$$

but the data actually follows

$$y_t = g(x_t,\gamma_n^\circ) + e_t \qquad t = 1, 2, \ldots, n$$

where

$$g(x,\gamma) = \Sigma_{j=0}^{\infty}\gamma_j x^j .$$

The equality is with respect to uniform convergence. That is, one restricts attention to the set Γ^* of $\gamma = (\gamma_0, \gamma_1, \ldots)$ with

$$\ell im_{J\to\infty} \sup_{x\in[0,1]} |\Sigma_{j=0}^{J}\gamma_j x^j| < \infty$$

and $g(x,\gamma)$ denotes that continuous function on $[0,1]$ with

$$\ell im_{J\to\infty} \sup_{x\in[0,1]} |g(x,\gamma) - \Sigma_{j=0}^{J}\gamma_j x^j| = 0 .$$

Take γ as equivalent to γ° and write $\gamma = \gamma^\circ$ if $g(x,\gamma) = g(x,\gamma^\circ)$ for all x in $[0,1]$. Define

$$\rho(\gamma,\gamma^\circ) = \ell im_{J\to\infty} \sup_{x\in[0,1]} |\Sigma_{j=0}^{J}(\gamma_j - \gamma_j^\circ)x^j| .$$

Show that (Γ^*,ρ) is a metric space on these equivalence classes (Royden, 1963, Section 7.1). If the model is fitted by least squares, if $f(x,\lambda)$ is continuous over $[0,1]\times\Lambda$, and if the estimation space Λ is compact we will show later that

$$\lambda_n^\circ \text{ minimizes } s_n^\circ(\lambda) = (1/n)\Sigma_{t=1}^n[g(x_t,\gamma_n^\circ) - f(x_t,\lambda)]^2 .$$

Assume that $f(x,\lambda)$ and $\{x_t\}$ are such that $s_n^\circ(\lambda)$ has a unique minimum for n larger than some N . By the Stone-Weirstrass Theorem (Royden, 1963, Section 9.6) we can find a γ° in Γ^* with

$$g(x,\gamma^\circ) = f(x,\lambda^* + \Delta/\sqrt{n}) .$$

That is, $\ell\text{im}_{J\to\infty}\sup_{x\in[0,1]}|\Sigma_{j=0}^J\gamma_j^\circ x^j - f(x,\lambda^*+\Delta/\sqrt{n})| = 0$.

Show that it is possible to truncate γ° at some point m_n such that if

$$\gamma_n^\circ = (\gamma_0, \gamma_1, \ldots, \gamma_{m_n}, 0, \ldots)$$

then

$$|\lambda_n^\circ - \lambda^* - \Delta/\sqrt{n} \ | < 1/n$$

for $n > N$. Hint: See the proof of Theorem 3. Show that

$$\ell\text{im}_{n\to\infty}\rho(\gamma_n^\circ,\gamma^*) = 0 .$$

Let $\Gamma = \{\gamma_n^\circ\}_{n=N}^\infty$. Show that (Γ,ρ) is a compact metric space.

3. (Construction of a Pitman drift). Let g(x) be once continuously differentiable on a bounded, open, convex set in R^k containing $\mathcal{X}$. By rescaling the data,we may assume that $\mathcal{X} \subset X_{i=1}^k[0,2\pi]$ without loss of generality. Then g(x) can be expanded in a multivariate Fourier series.

Letting r denote a multi-index, a multi-index being a vector with integer (positive, negative, or zero) components and letting $|r| = \Sigma_{i=1}^{k} |r_i|$, a multivariate Fourier series of order R is written $\Sigma_{|r| \le R} \gamma_r e^{ir'x}$ with $e^{ir'x} = \cos(r'x) + i \sin(r'x)$ and $i = \sqrt{-1}$. The restriction $\gamma_r = \bar{\gamma}_{-r}$ where the overbar denotes complex conjugation will cause the Fourier series to be real valued. We have (Edmunds and Moscatelli, 1977)

$$\ell im_{R \to \infty} \sup_{\mathcal{X}} |g(x) - \Sigma_{|r| \le R} \gamma_r \, e^{ir'x}| = 0 \, .$$

Construct a Pitman drift using a multivariate Fourier expansion along the same lines as in Problem 2.

4. Show that if for any $\epsilon > 0$ there is a $\delta > 0$ such that $|v - v^{\circ}| < \delta$ implies that

$$h(v^{\circ},\alpha) - \epsilon < h(v,\alpha) < h(v^{\circ},\alpha) + \epsilon$$

for all α in Θ then $\sup_{\alpha \in \Theta} h(v,\alpha)$ and $\inf_{\alpha \in \Theta} h(v,\alpha)$ are continuous.

5. Referring to the proof of Theorem 2 , show that $\{\ell' f[Y(e,x,\gamma),x,\tau,\lambda]\}^2 \le b(e,x)$ implies that

$$\{\int_{\mathcal{E}} \ell' f[Y(e,x,\gamma),x,\tau,\lambda] \, dP(e)\}^2 \le \int_{\mathcal{E}} \{\ell' f[Y(e,x,\gamma),x,\tau,\lambda]\}^2 \, dP(e) \le b(x) \, .$$

Show that $\ell im_{n \to \infty} (1/n) V_n = \ell' I^* \ell$.

6. Show that if $\hat{\tau}_n$ converges almost surely to τ^* and $\sqrt{n}(\hat{\tau}_n - \tau_n^{\circ})$ is bounded in probability then $\ell im_{n \to \infty} \tau_n^{\circ} = \tau^*$.

3. LEAST MEAN DISTANCE ESTIMATORS

Recall that a least mean distance estimator $\hat{\lambda}_n$ is defined as the solution of the optimization problem

$$\text{Minimize:} \quad s_n(\lambda) = (1/n)\Sigma_{t=1}^{n} s(y_t,x_t,\hat{\tau}_n,\lambda)$$

where $\hat{\tau}_n$ is a random variable which corresponds conceptually to estimators of nuisance parameters. A constrained least mean distance estimator $\tilde{\lambda}_n$ is the solution of the optimization problem

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

where $h(\lambda)$ maps R^p into R^q .

The objective of this section is to find the almost sure limit and the asymptotic distribution of the unconstrained estimator $\hat{\lambda}_n$ under regularity conditions that do not rule out specification error. Some ancillary facts regarding the asymptotic distribution of the constrained estimator $\tilde{\lambda}_n$ under a Pitman drift are also derived for use in later sections on hypothesis testing. In order to permit this Pitman drift, and to allow generality that may be useful in other contexts, the parameter γ_n^o of the data generating model is permitted to depend on the sample size n throughout. A more conventional asymptotic theory regarding the unconstrained estimator $\hat{\lambda}_n$ is obtained by applying these results with γ_n^o held fixed at a point γ^* for all n . These results are due to Souza (1979) in the main with some refinements made here to center $\hat{\lambda}_n$ about a point λ_n^o so as to isolate results regarding $\hat{\lambda}_n$ from the Pitman drift assumption.

An example, a correctly specified iteratively rescaled M-estimator, is carried throughout the discussion to illustrate how the regularity conditions may be satisfied in correctly specified situations.

EXAMPLE 1. (Iteratively rescaled M-estimator) The data generating model is

$$y_t = f(x_t, \gamma_n^{\circ}) + e_t \qquad t = 1, 2, \ldots, n \; .$$

An estimate of scale is obtained by first minimizing

$$(1/n) \, \Sigma_{t=1}^{n} \, \rho[y_t - f(x_t, \theta)]$$

with respect to θ to obtain $\hat{\theta}_n$ where

$$\rho(u) = \ell n \cosh(u/2)$$

and then solving

$$(1/n) \, \Sigma_{t=1}^{n} \Psi^2\{[y_t - f(x_t, \hat{\theta}_n)]/\tau\} - \int \Psi^2(e) \, d\Phi(e)$$

with respect to τ to obtain $\hat{\tau}_n$ where

$$\Psi(u) = (d/du)\rho(u) = \tfrac{1}{2} \tanh(u/2)$$

and Φ is the standard normal distribution function. The parameters of the model are estimated by minimizing

$$s_n(\lambda) = (1/n) \Sigma_{t=1}^{n} \, \rho\{[y_t - f(x_t, \lambda)]/\hat{\tau}_n\} \; ,$$

whence

$$s(y, x, \tau, \lambda) = \rho\{[y - f(x, \lambda)]/\tau\} \; .$$

The error distribution P(e) is symmetric, puts positive probability on every open interval of the real line and has finite first and second moments. See Huber (1964) for the motivation. □

The first question one must address is: What is $\hat{\lambda}_n$ to be regarded as estimating in a finite sample. Ordinarily, in an asymptotic estimation theory, the parameter γ^o of the data generating model is held fixed and $\hat{\lambda}_n$ would be regarded as estimating the almost sure limit λ^* of $\hat{\lambda}_n$ in each finite sample. But we have both misspecification and a parameter γ_n^o that is subject to drift and either of these situations is enough to make that answer to the question unsatisfactory. If we regarded $\hat{\lambda}_n$ as centered about its almost sure limit λ^* (Theorem 3), we would find it necessary to impose a Pitman drift, accelerate the rate of convergence of Cesaro sums generated from $\{x_t\}_{t=1}^{\infty}$, or impose other regularity conditions to show that $\sqrt{n}(\hat{\lambda}_n - \lambda^*)$ is asymptotically normally distributed. Such conditions are unnatural in an estimation setting. A more satisfactory answer to the question is obtained if one regards $\hat{\lambda}_n$ as estimating λ_n^o that is the solution to

$$\text{Minimize:} \quad s_n^o(\lambda) = (1/n)\Sigma_{t=1}^n \int_{\mathcal{E}} s[Y(e,x_t,\gamma_n^o),x_t,\tau_n^o,\lambda]\, dP(e) \ ;$$

τ_n^o is defined later (Assumption 4). With this choice, one can show that $\sqrt{n}(\hat{\lambda}_n - \lambda_n^o)$ is asymptotically normally distributed without unusual regularity conditions. Moreover, in analytically tractable situations such as a linear model fitted by least squares to data that actually follow a nonlinear model, it turns out that λ_n^o is indeed the mean of $\hat{\lambda}_n$ in finite samples.

We call the reader's attention to some heavily used notation and then state the identification condition:

NOTATION 1.

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

$s_n^o(\lambda) = (1/n)\Sigma_{t=1}^{n} \int_{\mathcal{E}} s[Y(e, x_t, \gamma_n^o), x_t, \tau_n^o, \lambda]\ dP(e)$

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

$\hat{\lambda}_n$ minimizes $s_n(\lambda)$

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

λ_n^o minimizes $s_n^o(\lambda)$

λ_n^* minimizes $s_n^o(\lambda)$ subject to $h(\lambda) = 0$

λ^* minimizes $s^*(\lambda)$

ASSUMPTION 4. (Identification) The parameter γ° is indexed by n and the sequence $\{\gamma_n^{\circ}\}$ converges to a point γ^*. The sequence of nuisance parameter estimators is centered at a point τ_n° in the sense that $\sqrt{n}(\hat{\tau}_n - \tau_n^{\circ})$ is bounded in probability; the sequence $\{\tau_n^{\circ}\}$ converges to a point τ^* and $\{\hat{\tau}_n\}$ converges almost surely to τ^*. $s^*(\lambda)$ has a unique minimum over the estimation space Λ^* at λ^*.

The critical condition imposed by Assumption 4 is that $s^*(\lambda)$ must have a unique minimum over Λ^*. In a correctly specified situation, the usual approach to verification is to commence with an obviously minimal identification condition. Then known results for the simple location problem that motivated the choice of distance function $s(y,x,\tau,\lambda)$ are exploited to verify a unique association of λ^* to γ^* over Λ^*. We illustrate with the example:

EXAMPLE 1. (Continued) We are trapped in a bit of a circularity in that we need the results of this section and the next in order to compute the center τ_n^o of the nuisance parameter estimator $\hat{\tau}_n$ and to show that $\sqrt{n}(\hat{\tau}_n - \tau_n^o)$ is bounded in probability. So we must defer verification until the end of Section 4. At that time we shall find that $\tau_n^o, \tau^* > 0$ which fact we shall use now.

To verify that $s^*(\lambda)$ has a unique minimum one first notes that it will be impossible to determine λ by observing $\{y_t, x_t\}$ if $f(x,\lambda) = f(x,\gamma)$ for $\lambda \neq \gamma$ at each x in $\mathcal{X}$ that is given weight by the measure μ . Then a minimal identification condition is

$$\lambda \neq \gamma \Rightarrow \mu\{x: f(x,\lambda) \neq f(x,\gamma)\} > 0 .$$

This is a condition both on the function $f(x,\lambda)$ and the infinite sequence $\{x_t\}_{t=1}^{\infty}$.

Now for $\tau > 0$

$$\varphi(\delta) = \int_{\mathcal{E}} \rho[(e+\delta)/\tau] \; dP(e)$$

is known (Problem 9) to have a unique minimum at $\delta = 0$ when $P(e)$ is symmetric about zero, has finite first moment, and assigns positive probability to every nonempty, open interval. Let

$$\delta(x) = f(x,\gamma) - f(x,\lambda) .$$

If $\lambda \neq \gamma$ then $\varphi[\delta(x)] \geq \varphi(0)$ for every x . Again, if $\lambda \neq \gamma$ the identification condition implies that $\varphi[\delta(x)] > \varphi(0)$ on some set A of positive μ measure. Consequently, if $\lambda \neq \gamma$

$$s^*(\gamma,\tau,\lambda) = \int_{\mathcal{X}} \varphi[\delta(x)]\, d\mu(x) > \int_{\mathcal{X}} \varphi(0)\, d\mu(x) = \varphi(0) \ .$$

Now $s^*(\lambda) = s^*(\gamma^*,\tau^*,\lambda)$ so that $s^*(\lambda) > \varphi(0)$ if $\lambda \neq \gamma^*$ and $s^*(\lambda) = \varphi(0)$ if $\lambda = \gamma^*$ which shows that $s^*(\lambda)$ has a unique minimum at $\lambda = \gamma^*$.

A similar argument can be used to compute λ_n°. It runs as follows. Let

$$s_n^\circ(\gamma,\tau,\lambda) = (1/n)\Sigma_{t=1}^n \varphi[\delta(x_t)] \geq (1/n)\Sigma_{t=1}^n \varphi(0) = \varphi(0) \ .$$

Since $s_n^\circ(\lambda) = s_n^\circ(\gamma_n^\circ,\tau_n^\circ,\lambda)$, $s_n^\circ(\lambda)$ has a minimum at $\lambda = \gamma_n^\circ$. It is not necessary to the theory which follows that λ_n° be unique. Existence is all that is required. Similarly for τ_n° . $\square$

We shall adjoin some technical conditions. To comment, note that 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 (Lemma 2) to $\hat{\tau}_n$. Consequently, without loss of generality, it may be assumed that $\hat{\tau}_n$ takes its values in a compact ball T for which τ^* is an interior point. Thus, the effective conditions of the next assumption are domination of the objective function and a compact estimation space Λ^*. As noted in the previous section, a compact estimation space is not a serious restriction in applications.

ASSUMPTION 5. The estimation space Λ^* is compact; $\{\hat{\tau}_n\}$ and $\{\tau_n^\circ\}$ are contained in T which is a closed ball centered at τ^* with finite, nonzero radius. The distance function $s(y,x,\tau,\lambda)$ is continuous on $\mathcal{Y} \times \mathcal{X} \times T \times \Lambda^*$ and $|s(y,x,\tau,\lambda)| \leq b[q(y,x,\gamma),x]$ on $\mathcal{Y} \times \mathcal{X} \times T \times \Lambda^* \times \Gamma$; $b(e,x)$ is that of Assumption 3.

The exhibition of the requisite dominating function b(e,x) is an ad hoc process and one exploits the special characteristics of an application. We illustrate with Example 1:

EXAMPLE 1. (Continued) Now $\rho(u) \leq (1/2)|u|$ (Problem 9) so that

$$\begin{aligned} |s(y,x,\tau,\lambda)| &= \rho\{[e + f(x,\gamma) - f(x,\lambda)]/\tau\} \\ &\leq \tfrac{1}{2}|e + f(x,\gamma) - f(x,\lambda)|/\tau \\ &\leq [\,|\,e\,| + \sup_{\Gamma}|f(x,\gamma)| + \sup_{\Lambda^*}|f(x,\lambda)|\,]/\min T\ . \end{aligned}$$

Suppose that $\Gamma = \Lambda^*$ and that $\sup_{\Gamma} |f(x,\gamma)|$ is μ-integrable. Then

$$b_1(e,x) =[\,|\,e\,| + 2 \sup_{\Gamma}|f(x,\gamma)|\,]/\min T$$

will serve to dominate $s(y,x,\tau,\lambda)$. If $\mathcal{X}$ is compact then $b_1(e,x)$ is integrable for any μ . To see this observe that $f(x,\gamma)$ must be continuous over $\mathcal{X} \times \Gamma$ to satisfy Assumption 2. A continuous function over a compact set is bounded so $\sup_{\Gamma} f(x,\gamma)$ is a bounded, measurable function.

Later (Assumption 6) we shall need to dominate

$$\begin{aligned} &\|(\partial/\partial\lambda)s(y,x,\tau,\lambda)\| \\ &\quad = \|\Psi\{[e + f(x,\gamma) - f(x,\lambda)]/\tau\}(\partial/\partial\lambda)f(x,\lambda)/\tau\| \\ &\quad \leq \sup_{\Gamma}\|(\partial/\partial\lambda)\ f(x,\lambda)\|/\min T \end{aligned}$$

since $|\Psi(u)| = |(1/2)\tanh(u/2)| \leq (1/2)$. Thus

$$b_2(e,x) = \sup_{\Gamma}\ \|(\partial/\partial\lambda)\ f(x,\lambda)\|/\min T$$

serves as a dominating function.

One continues the construction of suitable $b_1(e,x)$, $b_2(e,x)$, ... to dominate each of the functions listed in Assumptions 4 and 6 . Then the overall dominating function of Assumption 3 is

$$b(e,x) = \Sigma_i b_i(x,e)\ .$$

This construction will satisfy the formal logical requirements of the theory. In many applications $\mathfrak{X}$ can be taken as compact and P(e) to possess enough moments so that the domination requirements of the general theory obtain trivially. $\square$

We can now prove that $\hat{\lambda}_n$ is a strongly consistent estimator of λ^*; first a lemma, then the proof:

LEMMA 1. Let Assumptions 1 through 5 hold. Then $s_n(\lambda)$ converges almost surely to $s^*(\lambda)$ uniformly on Λ^* and $s_n^o(\lambda)$ converges to $s^*(\lambda)$ uniformly on Λ^*.

PROOF. We shall prove the result for $s_n(\lambda)$. The argument for $s_n^o(\lambda)$ is much the same (Problem 1). Now

$$\begin{aligned}
&\sup_{\Lambda^*}|s_n(\lambda) - s^*(\lambda)| \\
&\quad\le \sup_{\Lambda^*}|(1/n)\Sigma_{t=1}^n s[Y(e_t,x_t,\gamma_n^o),x_t,\hat{\tau}_n,\lambda] \\
&\qquad\qquad - \int_{\mathcal{X}}\int_{\mathcal{E}} s[Y(e,x,\gamma_n^o),x,\hat{\tau}_n,\lambda]\, dP(e)\, d\mu(x)| \\
&\quad + \sup_{\Lambda^*}|\int_{\mathcal{X}}\int_{\mathcal{E}} s[Y(e,x,\gamma_n^o),x,\hat{\tau}_n,\lambda]\, dP(e)\, d\mu(x) \\
&\qquad\qquad - \int_{\mathcal{X}}\int_{\mathcal{E}} s[Y(e,x,\gamma^*),x,\tau^*,\lambda]\, dP(e)\, d\mu(x)| \\
&\quad\le \sup_{\Gamma\times T\times\Lambda^*}|(1/n)\Sigma_{t=1}^n\, s[Y(e_t,x_t,\gamma),x_t,\tau,\lambda] \\
&\qquad\qquad - \int_{\mathcal{X}}\int_{\mathcal{E}} s[Y(e,x,\gamma),x,\tau,\lambda]\, dP(e)\, d\mu(x)| \\
&\quad + \sup_{\Lambda^*}\int_{\mathcal{X}}\int_{\mathcal{E}} |s[Y(e,x,\gamma_n^o),x,\hat{\tau}_n,\lambda] \\
&\qquad\qquad - s[Y(e,x,\gamma^*),x,\tau^*,\lambda]|dP(e)\, d\mu(x) \\
&\quad = \sup_{\Gamma\times T\times\Lambda^*} f_n(\gamma,\tau,\lambda) + \sup_{\Lambda^*} g(\gamma_n^o,\hat{\tau}_n,\lambda)\ .
\end{aligned}$$

Since $\Gamma \times T \times \Lambda^*$ is compact, and $s(y,x,\tau,\lambda)$ is continuous on $\mathcal{Y} \times \mathcal{X} \times T \times \Lambda^*$ with $|s(y,x,\tau,\lambda)| \le b[q(y,x,\gamma),x]$ for all (y,x) in $\mathcal{Y} \times \mathcal{X}$ and all (γ,τ,λ) in $\Gamma \times T \times \Lambda^*$ we have, by Theorem 1, that $\sup_{\Gamma\times T\times\Lambda^*} f_n(\gamma,\tau,\lambda)$ converges almost surely to zero. Given any sequence $\{(\gamma_n,\tau_n,\lambda_n)\}$ that converges to, say, $(\gamma^o,\tau^o,\lambda^o)$ we have, by the Dominated Convergence Theorem with $2b(e,x)$ being the dominating function, that $\ell im_{n\to\infty} g(\gamma_n,\tau_n,\lambda_n) = g(\gamma^o,\tau^o,\lambda^o)$. This shows that $g(\gamma,\tau,\lambda)$ is continuous in (γ,τ,λ). Moreover, $\sup_{\Lambda^*} g(\gamma,\tau,\lambda)$ is continuous in (γ,τ) ; see the proof of Theorem 1 for details. Then, since $(\gamma_n^o,\hat{\tau}_n)$ converges almost surely to (γ^*,τ^*), $\sup_{\Lambda^*} g(\gamma_n^o,\hat{\tau}_n,\lambda)$ converges almost surely to zero. ▯

THEOREM 3. (Strong consistency) Let Assumptions 1 through 5 hold. Then $\hat{\lambda}_n$ converges almost surely to λ^* and λ_n° converges to λ^*.

PROOF. If a realization $\{e_t\}$ of the errors is held fixed then $\{\hat{\lambda}_n\}$ becomes a fixed, vector-valued sequence and $\{s_n(\lambda)\}$ becomes a fixed sequence of functions. We shall hold fixed a realization $\{e_t\}$ with the attribute that $s_n(\lambda)$ converges uniformly to $s^*(\lambda)$ on Λ^*; almost every realization is such by Lemma 1. If we can show that the corresponding sequence $\{\hat{\lambda}_n\}$ converges to λ^* then we have the first result. This is the plan.

Now $\hat{\lambda}_n$ lies in the compact set Λ^*. Thus the sequence $\{\hat{\lambda}_n\}$ has at least one limit point $\hat{\lambda}$ and one subsequence $\{\hat{\lambda}_{n_m}\}$ with $\ell im_{m\to\infty}\hat{\lambda}_{n_m} = \hat{\lambda}$. Now, by uniform convergence (see Problem 2),

$$\begin{aligned} s^*(\hat{\lambda}) &= \ell im_{m\to\infty} s_{n_m}(\hat{\lambda}_{n_m}) \\ &\leq \ell im_{m\to\infty} s_{n_m}(\lambda^*) \\ &= s^*(\lambda^*) \end{aligned}$$

where the inequality is due to the fact that $s_n(\hat{\lambda}_n) \leq s_n(\lambda^*)$ for every n as $\hat{\lambda}_n$ is a minimizing value. The assumption of a unique minimum, Assumption 4, implies $\hat{\lambda} = \lambda^*$. Then $\{\hat{\lambda}_n\}$ has only the one limit point λ^*.

An analogous argument implies that λ_n° converges to λ^* (Problem 3). ▯

The following notation defines the parameters of the asymptotic distribution of $\hat{\lambda}_n$.

NOTATION 2.

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

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

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

$$\mathfrak{J}^* = \bar{\mathfrak{J}}(\lambda^*)\ ,\ \mathcal{J}^* = \bar{\mathcal{J}}(\lambda^*)\ ,\ \mathcal{U}^* = \bar{\mathcal{U}}(\lambda^*)$$

If this were maximum likelihood estimation with $s(y,x,\tau,\lambda) = -\ell n\, p(y|x,\lambda)$, then $\mathcal{I}^*$ would be the information matrix and $\mathcal{J}^*$ the expectation of the Hessian of the log likelihood. Under correct specification one would have $\mathcal{U}^* = 0$ and $\mathcal{I}^* = \mathcal{J}^*$ (Section 7).

We illustrate the computations with the example.

EXAMPLE 1. (Continued) The first and second derivatives of $s(y,x,\tau,\lambda)$ are:

$$
\begin{aligned}
(\partial/\partial\lambda)s(y,x,\tau,\lambda) &= (\partial/\partial\lambda)\rho\ \{[y-f(x,\lambda)]/\tau\} \\
&= (-1/\tau)\Psi\{[y-f(x,\lambda)]/\tau\}(\partial/\partial\lambda)\ f(x,\lambda)
\end{aligned}
$$

$$
\begin{aligned}
(\partial^2/\partial\lambda\partial\lambda')s(y,x,\tau,\lambda) &= (\partial/\partial\lambda)(-1/\tau)\Psi\{[y-f(x,\lambda)]/\tau\}(\partial/\partial\lambda')f(x,\lambda) \\
&= (1/\tau^2)\Psi'\{[y-f(x,\lambda)]/\tau\}[(\partial/\partial\lambda)f(x,\lambda)][(\partial/\partial\lambda)f(x,\lambda)]' \\
&\quad - (1/\tau)\Psi\{[y-f(x,\lambda)]/\tau\}(\partial^2/\partial\lambda\partial\lambda')f(x,\lambda)\ .
\end{aligned}
$$

Evaluating the first derivative at $y = f(x,\gamma) + e$, $\tau = \tau^*$, and $\lambda = \gamma$ we have

$$
\begin{aligned}
&\int_{\mathcal{E}}(\partial/\partial\lambda)s[Y(e,x,\gamma),x,\tau^*,\lambda]\ dP(e)\Big|_{\gamma=\lambda} \\
&\quad = (-1/\tau^*)\int_{\mathcal{E}}\Psi(e/\tau^*)\ dP(e)\ (\partial/\partial\lambda)\ f(x,\lambda) \\
&\quad = (-1/\tau^*)(0)(\partial/\partial\lambda)\ f(x,\lambda) \\
&\quad = 0
\end{aligned}
$$

because $\Psi(e/\tau)$ is an odd function, that is $\Psi(u) = \Psi(-u)$, and an odd function integrates to zero against a symmetric error distribution. Thus, $u^* = 0$. In fact, u^* is always zero in a correctly specified situation when using a sensible estimation procedure. To continue, writing $\mathcal{E}\Psi^2(e/\tau^*)$ for $\int_{\mathcal{E}}\Psi^2(e/\tau^*)\,dP(e)$ and $\mathcal{E}\Psi'(e/\tau^*)$ for $\int_{\mathcal{E}}(d/du)\Psi(u)\Big|_{u=e/\tau^*}dP(e)$, we have

$$
\begin{aligned}
&\int_{\mathcal{E}}\{(\partial/\partial\lambda)s[Y(e,x,\gamma),x,\tau^*,\lambda]\}\{(\partial/\partial\lambda)s[Y(e,x,\gamma),x,\tau^*,\lambda]\}'dP(e)\Big|_{\gamma=\lambda} \\
&\quad = (1/\tau^*)^2\ \mathcal{E}\Psi^2(e/\tau^*)[(\partial/\partial\lambda)f(x,\lambda)][(\partial/\partial\lambda)f(x,\lambda)]'
\end{aligned}
$$

and

$$\int_{\mathcal{E}} (\partial^2/\partial\lambda\partial\lambda') s[Y(e,x,\gamma),x,\tau^*,\lambda]\ dP(e)\Big|_{\gamma = \lambda}$$

$$= (1/\tau^*)^2\ \mathcal{E}\Psi'(e/\tau^*)[(\partial/\partial\lambda)\ f(x,\lambda)][(\partial/\partial\lambda)\ f(x,\lambda)]' \quad .$$

Thus,

$$\mathcal{I}^* = (1/\tau^*)^2\mathcal{E}\ \Psi^2(e/\tau^*)\int_{\mathcal{X}}[(\partial/\partial\lambda)\ f(x,\lambda^*)][(\partial/\partial\lambda)f(x,\lambda^*)]'\ d\mu(x)$$

and

$$\mathcal{J}^* = (1/\tau^*)^2\mathcal{E}\Psi'(e/\tau^*)\int_{\mathcal{X}}[(\partial/\partial\lambda)f(x,\lambda^*)][(\partial/\partial\lambda)f(x,\lambda^*)]'\ d\mu(x)\ . \quad \square$$

In Section 5, the distributions of test statistics are characterized in terms of the following quantities:

NOTATION 3.

$$\bar{u}_n(\lambda) = (1/n)\Sigma_{t=1}^{n} \{\int_{\mathcal{E}} (\partial/\partial\lambda)s[Y(e,x_t,\gamma_n^\circ),x_t,\tau_n^\circ,\lambda]\, dP(e)\}$$

$$\times \{\int_{\mathcal{E}} (\partial/\partial\lambda)s[Y(e,x_t,\gamma_n^\circ),x_t,\tau_n^\circ,\lambda]\, dP(e)\}'$$

$$\bar{\mathscr{I}}_n(\lambda) = (1/n)\Sigma_{t=1}^{n}\int_{\mathcal{E}}\{(\partial/\partial\lambda)s[Y(e,x_t,\gamma_n^\circ),x_t,\tau_n^\circ,\lambda]\}$$

$$\times \{(\partial/\partial\lambda)s[Y(e,x_t,\gamma_n^\circ),x_t,\tau_n^\circ,\lambda]\}'dP(e) - \bar{u}_n(\lambda)$$

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

$$\mathscr{I}_n^\circ = \bar{\mathscr{I}}_n(\lambda_n^\circ) \;,\; \mathscr{J}_n^\circ = \bar{\mathscr{J}}_n(\lambda_n^\circ),\; u_n^\circ = \bar{u}_n(\lambda_n^\circ)$$

$$\mathscr{I}_n^* = \bar{\mathscr{I}}_n(\lambda_n^*) \;,\; \mathscr{J}_n^* = \bar{\mathscr{J}}_n(\lambda_n^*),\; u_n^* = \bar{u}_n(\lambda_n^*)$$

We illustrate their computation with Example 1:

EXAMPLE 1. (Continued) Let

$$\mu_t(\lambda) = \int_{\mathcal{E}} \Psi\{[e + f(x,\gamma_n^\circ) - f(x,\lambda)]/\tau_n^\circ\}\ dP(e)\ ,$$

$$\sigma_t^2(\lambda) = \int_{\mathcal{E}} \Psi^2\{[e + f(x,\gamma_n^\circ) - f(x,\lambda)]/\tau_n^\circ\}\ dP(e) - \mu_t^2(\lambda)\ ,$$

$$\beta_t(\lambda) = \int_{\mathcal{E}} \Psi'\{[e + f(x,\gamma_n^\circ) - f(x,\lambda)]/\tau_n^\circ\}\ dP(e)\ .$$

Note that if one evaluates at $\lambda = \lambda_n^\circ$ then $\mu_t(\lambda_n^\circ) = 0$, $\sigma_t^2(\lambda_n^\circ) = \mathcal{E}\ \Psi^2(e/\tau_n^\circ)$, and $\beta_t(\lambda_n^\circ) = \mathcal{E}\ \Psi'(e/\tau_n^\circ)$ which eliminates the variation with t, but, if one evaluates at $\lambda = \lambda_n^*$ then the variation with t remains. We have by direct computation that

$$\mathfrak{U}(\lambda) = (1/\tau_n^\circ)^2(1/n)\ \Sigma_{t=1}^n u_t^2(\lambda)\ [(\partial/\partial\lambda)f(x_t,\lambda)][(\partial/\partial\lambda)f(x_t,\lambda)]'\ ,$$

$$\mathfrak{J}(\lambda) = (1/\tau_n^\circ)^2(1/n)\ \Sigma_{t=1}^n \sigma_t^2(\lambda)\ [(\partial/\partial\lambda)f(x_t,\lambda)][(\partial/\partial\lambda)f(x_t,\lambda)]'\ ,$$

$$\mathscr{J}(\lambda) = (1/\tau_n^\circ)^2(1/n)\ \Sigma_{t=1}^n\ \beta_t(\lambda)[(\partial/\partial\lambda)f(x_t,\lambda)][(\partial/\partial\lambda)f(x_t,\lambda)]'$$

$$-\ (1/\tau_n^\circ)(1/n)\ \Sigma_{t=1}^n \mu_t(\lambda)\ (\partial^2/\partial\lambda\partial\lambda')f(x_t,\lambda)\ .\quad \square$$

Some plausible estimators of $\mathcal{I}^*$ and $\mathcal{J}^*$ --- or of $(\mathcal{I}_n^\circ, \mathcal{J}_n^\circ)$ and $(\mathcal{I}_n^*, \mathcal{J}_n^*)$ respectively depending on one's point of view --- are as follows:

NOTATION 4.

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

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

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

We illustrate the computations and point out some alternatives using Example 1.

EXAMPLE 1. (Continued) Let

$$\hat\Psi_t = \Psi\{[y_t - f(x_t,\hat\lambda_n)]/\hat\tau_n\}$$

$$\hat\Psi'_t = \Psi'\{[y_t - f(x_t,\hat\lambda_n)]/\hat\tau_n\}$$

then

$$\hat{\mathscr{I}}_n = (1/\hat\tau_n)^2(1/n)\Sigma_{t=1}^n\hat\Psi_t^2[(\partial/\partial\lambda)f(x_t,\hat\lambda_n)][(\partial/\partial\lambda)f(x_t,\hat\lambda_n)]'$$

$$\hat{\mathscr{J}}_n = (1/\hat\tau_n)^2(1/n)\Sigma_{t=1}^n\hat\Psi'_t[(\partial/\partial\lambda)f(x_t,\hat\lambda_n)][(\partial/\partial\lambda)f(x_t,\hat\lambda_n)]' - (1/\hat\tau_n)(1/n)\Sigma_{t=1}^n\hat\Psi_t(\partial^2/\partial\lambda\partial\lambda')f(x_t,\hat\lambda_n) \ .$$

Alternatives that are similar to the forms used in least squares are

$$\hat I_n = (\hat\sigma_n^2/\hat\tau_n^2)\ (1/n)\Sigma_{t=1}^n[(\partial/\partial\lambda)f(x_t,\hat\lambda_n)][(\partial/\partial\lambda)f(x_t,\hat\lambda_n)]'$$

$$\hat J_n = (\hat\beta_n/\hat\tau_n^2)\ (1/n)\Sigma_{t=1}^n[(\partial/\partial\lambda)f(x_t,\hat\lambda_n)][(\partial/\partial\lambda)f(x_t,\hat\lambda_n)]'$$

with

$$\hat\sigma_n^2 = (1/n)\Sigma_{t=1}^n\hat\Psi_t^2$$

$$\hat\beta_n = (1/n)\Sigma_{t=1}^n\hat\Psi'_t \ . \quad \square$$

Some additional, technical restrictions needed to prove asymptotic normality are:

ASSUMPTION 6. The parameter space Λ^* contains a closed ball Λ centered at λ^* with finite, nonzero radius such that the elements of $(\partial/\partial\lambda)s(y,x,\tau,\lambda)$, $(\partial^2/\partial\lambda\partial\lambda')s(y,x,\tau,\lambda)$, $(\partial^2/\partial\tau\partial\lambda')s(y,x,\tau,\lambda)$, and $[(\partial/\partial\lambda)s(y,x,\tau,\lambda)][(\partial/\partial\lambda)s(y,x,\tau,\lambda)]'$ are continuous and dominated by $b[q(y,x,\gamma),x]$ on $\mathcal{Y} \times \mathcal{X} \times \Gamma \times T \times \Lambda$. Moreover, $\mathcal{J}^*$ is nonsingular and

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

The integral condition is sometimes encountered in the theory of maximum likelihood estimation; see Durbin (1970) for a detailed discussion. It validates the application of maximum likelihood theory to a subset of the parameters when the remainder are treated as if known in the derivations but are subsequently estimated. The assumption plays the same role here. It can be avoided in maximum likelihood estimation at a cost of additional complexity in the results; see Gallant and Holly (1980) for details. It can be avoided here as well but there is no reason to further complicate the results in view of the intended applications. In an application where the condition is not satisfied, the simplest solution is to estimate λ and τ jointly and not use a two-step estimator. We illustrate with the example:

EXAMPLE 1. (Continued)

$$(\partial^2/\partial\tau\partial\lambda')s[Y(e,x,\gamma^*),x,\tau,\lambda^*)]\Big|_{\gamma^*=\lambda^*} = (1/\tau^2)[\Psi(e/\tau)+\Psi'(e/\tau)(e/\tau)](\partial/\partial\lambda)f(x,\lambda^*)$$

Both $\Psi(e/\tau)$ and $\Psi'(e/\tau)(e/\tau)$ are odd functions and will integrate to zero for symmetric $P(e)$. ▯

The derivatives of the distance function plays the same role here as does the derivative of the log density function or score in maximum likelihood estimation. Hence,we use the same terminology here. As with the scores in maximum likelihood estimation, their normalized sum is asymptotically normally distributed:

THEOREM 4. (Asymptotic normality of the scores) Under Assumptions 1 through 6

$$\sqrt{n}\ (\partial/\partial\lambda)\ s_n(\lambda_n^\circ) \xrightarrow{\mathcal{L}} N(0,\mathcal{J}^*)\ .$$

$\mathcal{J}^*$ may be singular.

PROOF. By Theorem 2

$$(1/\sqrt{n})\Sigma_{t=1}^{n}\{(\partial/\partial\lambda)s(y_t,x_t,\hat{\tau}_n,\lambda_n^\circ) - \int_{\mathcal{E}}(\partial/\partial\lambda)s[Y(e,x_t,\gamma_n^\circ),x_t,\tau_n^\circ,\lambda_n^\circ]\,dP(e)\}\xrightarrow{\mathcal{L}}N(0,\mathcal{J}^*)\ .$$

Domination permits the interchange of differentiation and integration (Problem 11) and λ_n° is defined as a minimizing value whence

$$\Sigma_{t=1}^{n}\int_{\mathcal{E}}(\partial/\partial\lambda)s[Y(e,x_t,\gamma_n^\circ),x_t,\tau_n^\circ,\lambda_n^\circ]\,dP(e)$$

$$= \Sigma_{t=1}^{n}(\partial/\partial\lambda)\int_{\mathcal{E}}s[Y(e,x_t,\gamma_n^\circ),x_t,\tau_n^\circ,\lambda_n^\circ]\,dP(e)$$

$$= 0\ . \quad \square$$

We can now show that $\hat{\lambda}_n$ is asymptotically normally distributed. First we prove two lemmas:

LEMMA 2. (Tail equivalence). Let $\{\lambda_n\}$ be a sequence of vector-valued random variables that take their values in $\Lambda^* \subset R^p$ and that converge almost surely to a point λ^* in Λ^*. Let $\{s_n(\lambda)\}$ be a sequence of real valued random functions defined on Λ^*. Let $g(\lambda)$ be a vector-valued function defined on Λ^*. Let Λ° be an open subset of R^p with $\lambda^* \in \Lambda^\circ \subset \Lambda^*$. Then there is a sequence $\{\bar{\lambda}_n\}$ of random variables that take their values in Λ°, that satisfy

$$g(\lambda_n) = g(\bar{\lambda}_n) + o_s(n^{-\alpha})$$

for every $\alpha > 0$, and such that:

i) If $(\partial/\partial\lambda)s_n(\lambda)$ is continuous on Λ° and λ_n minimizes $s_n(\lambda)$ over Λ^* then

$$(\partial/\partial\lambda)s_n(\bar{\lambda}_n) = o_s(n^{-\alpha})$$

for every $\alpha > 0$.

ii) If $(\partial/\partial\lambda)s_n(\lambda)$ and $(\partial/\partial\lambda')h(\lambda)$ are continuous on Λ°, if $(\partial/\partial\lambda')h(\lambda)$ has full rank at $\lambda = \lambda^*$, and if λ_n minimizes $s_n(\lambda)$ over Λ^* subject to $h(\lambda) = 0$, then there is a vector $\bar{\theta}_n$ of (random) Lagrange multipliers such that

$$(\partial/\partial\lambda')[s_n(\bar{\lambda}_n) + \bar{\theta}_n' h(\bar{\lambda}_n)] = o_s(n^{-\alpha})$$

$$h(\bar{\lambda}_n) = o_s(n^{-\alpha})$$

for every $\alpha > 0$.

PROOF. The idea of the proof is that eventually λ_n is in Λ° and has, itself, the desired properties due to the almost sure convergence of λ_n to λ^*. Stating that the residual random variables are of almost sure order $o_s(n^{-\alpha})$ is just one way of expressing the fact that the requisite large n depends on the realization $\{e_t\}$ that obtains; that is, the convergence is not uniform in $\{e_t\}$.

We shall prove ii. By Problem 5 $(\partial/\partial\lambda')h(\lambda)$ has full rank on some open set $\mathcal{O}$ with $\lambda^* \in \mathcal{O} \subset \Lambda^\circ$. Define

$$\bar\lambda_n = \begin{cases} \lambda^* \text{ if } \lambda_n \notin \mathcal{O} \\ \lambda_n \text{ if } \lambda_n \in \mathcal{O} \end{cases} .$$

Fix a realization $\{e_t\}$ for which $\ell im_{n\to\infty}\lambda_n = \lambda^*$; almost every realization is such. There is an N such that $n > N$ implies $\lambda_n \in \mathcal{O}$ for all $n > N$. Since $\mathcal{O}$ is open and λ_n is the constrained optimum we have that $\bar\theta_n$ exists and that

$$\bar\lambda_n = \lambda_n$$

$$(\partial/\partial\lambda')[s_n(\lambda_n) + \bar\theta_n'h(\lambda_n)] = 0$$

$$h(\lambda_n) = 0$$

(Bartle, 1964, Sec. 21). Then, trivially,

$$\ell im_{n\to\infty} n^\alpha\|g(\bar\lambda_n) - g(\lambda_n)\| = 0$$

$$\ell im_{n\to\infty} n^\alpha\|(\partial/\partial\lambda')[s_n(\bar\lambda_n) + \bar\theta_n'h(\bar\lambda_n)]\| = 0,$$

$$\ell im_{n\to\infty} n^\alpha\|h(\bar\lambda_n)\| = 0 . \;\square$$

LEMMA 3. Under Assumptions 1 through 6, interchange of differentiation and integration is permitted in these instances:

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

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

$$(\partial/\partial\lambda)\ s_n^\circ(\lambda) = (1/n)\Sigma_{t=1}^n\int_{\mathcal{E}}(\partial/\partial\lambda)s[Y(e,x_t,\gamma_n^\circ),x_t,\tau_n^\circ,\lambda]\ dP(e),$$

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

Moreover:

$$\ell im_{n\to\infty}(\partial/\partial\lambda)s_n^\circ(\lambda) = (\partial/\partial\lambda)s^*(\lambda) \text{ uniformly on } \Lambda,$$

$$\ell im_{n\to\infty}(\partial^2/\partial\lambda\partial\lambda')s_n^\circ(\lambda) = (\partial^2/\partial\lambda\partial\lambda')s^*(\lambda) \text{ uniformly on } \Lambda,$$

$$\ell im_{n\to\infty}(\partial/\partial\lambda)s_n(\lambda) = (\partial/\partial\lambda)s^*(\lambda) \text{ almost surely, uniformly on } \Lambda,$$

$$\ell im_{n\to\infty}(\partial^2/\partial\lambda\partial\lambda')s_n(\lambda) = (\partial^2/\partial\lambda\partial\lambda')s^*(\lambda) \text{ almost surely, uniformly on } \Lambda\,.$$

PROOF. (Interchange) We shall prove the result for $(\partial/\partial\lambda)s^*(\lambda)$, the argument for the other three cases being much the same. Let λ in Λ and $\{h_m\}$ be any sequence with $\ell im_{n\to\infty}h_m = 0$ and $\lambda - h_m\xi_i$ in Λ where ξ_i is the $i^{\underline{th}}$ elementary vector. By the Mean Value Theorem,

$$\{s[Y(e,x,\gamma^*),x,\tau^*,\lambda] - s[Y(e,x,\gamma^*),x,\tau^*,\lambda - h_m\xi_i]\}/h_m$$
$$= (\partial/\partial\lambda_i)s[Y(e,x,\gamma^*),x,\tau^*,\lambda - \bar{h}_m(e,x)\xi_i]$$

where $|\bar{h}_m(e,x)| \le h_m$. (One can show that $\bar{h}_m(e,x)$ is measurable but it is not necessary for the validity of the proof as the composite function on the right hand side is measurable by virtue of being equal to the left.) Thus

$$[s^*(\lambda) - s^*(\lambda - h_m\xi_i)]/h_m$$
$$= \int_{\mathcal{X}}\int_{\mathcal{E}}(\partial/\partial\lambda_i)s[Y(e,x,\gamma^*),x,\tau^*,\lambda - \bar{h}_m(e,x)\xi_i]\ dP(e)\ d\mu(x)\ .$$

By the Dominated Convergence Theorem, with b(e,x) being the dominating function, and continuity

$$
\begin{aligned}
(\partial/\partial\lambda_i)s^*(\lambda) &= \ell im_{m\to\infty}[s^*(\lambda) - s^*(\lambda - h_m\xi_i)]/h_m \\
&= \int_{\mathcal{X}}\int_{\mathcal{E}} \ell im_{m\to\infty}(\partial/\partial\lambda_i)s[Y(e,x,\gamma^*),x,\tau^*,\lambda - \bar{h}_m(e,x)\xi_i]dP(e)\ d\mu(x) \\
&= \int_{\mathcal{X}}\int_{\mathcal{E}} (\partial/\partial\lambda_i)s[Y(e,x,\gamma^*),x,\tau^*,\lambda]dP(e)\ d\mu(x)\ .
\end{aligned}
$$

(Uniform convergence) The argument is the same as that used in the proof of Lemma 1. □

THEOREM 5. (Asymptotic normality) Let Assumptions 1 through 6 hold. Then:

$$\sqrt{n}(\hat{\lambda}_n - \lambda_n^o) \xrightarrow{\mathcal{L}} N[0, (\mathcal{J}^*)^{-1}\mathcal{I}^*(\mathcal{J}^*)^{-1}] ,$$

$\hat{\mathcal{I}}$ converges almost surely to $\mathcal{I}^* + \mathcal{U}^*$

$\mathcal{I}_n^o$ converges to $\mathcal{I}^*$,

$\mathcal{J}$ converges almost surely to $\mathcal{J}^*$,

$\mathcal{J}_n^o$ converges to $\mathcal{J}^*$.

PROOF. By Lemma 2, we may assume without loss of generality that $\hat{\lambda}_n$, $\lambda_n^o \in \Lambda$ and that $(\partial/\partial\lambda)s_n(\hat{\lambda}_n) = o_s(n^{-\frac{1}{2}})$, $(\partial/\partial\lambda)s_n^o(\lambda_n^o) = o(n^{-\frac{1}{2}})$, see Problem 6.

By Taylor's theorem

$$\sqrt{n}(\partial/\partial\lambda)s_n(\lambda_n^o) = \sqrt{n}(\partial/\partial\lambda)s_n(\hat{\lambda}_n) + \bar{\mathcal{J}}\sqrt{n}(\lambda_n^o - \hat{\lambda}_n)$$

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

$$(\partial/\partial\lambda')(\partial/\partial\lambda_i)s_n(\bar{\lambda}_{in})$$

and $\bar{\lambda}_{in}$ lies on the line segment joining λ_n^o to $\hat{\lambda}_n$. Now both λ_n^o and $\hat{\lambda}_n$ converge almost surely to λ^* by Theorem 3 so that $\bar{\lambda}_{in}$ converges almost surely to λ^*. Also, $(\partial/\partial\lambda')(\partial/\partial\lambda_i)s_n(\lambda)$ converges almost surely to $(\partial/\partial\lambda')(\partial/\partial\lambda_i)s^*(\lambda)$ uniformly on Λ by Lemma 3. Taking these two facts together (Problem 2), $\bar{\mathcal{J}}$ converges almost surely to $(\partial^2/\partial\lambda\partial\lambda')s(\lambda^*)$. By interchanging integration and differentiation as permitted by Lemma 3, $\mathcal{J}^* = (\partial^2/\partial\lambda\partial\lambda')s^*(\lambda^*)$. Thus we may write $\bar{\mathcal{J}} = \mathcal{J}^* + o_s(1)$ and as $(\partial/\partial\lambda)s_n(\hat{\lambda}_n) = o_s(n^{-\frac{1}{2}})$ we may write

$$[\mathcal{J}^* + o_s(1)]\sqrt{n}(\hat{\lambda}_n - \lambda_n^o) = -\sqrt{n}(\partial/\partial\lambda)s_n(\lambda_n^o) + o_s(1) .$$

The first result follows at once from Slutsky's theorem (Serfling, 1980, Sec. 1.5.4 or Rao, 1973, Sec. 2c.4).

By Theorem 1, with Assumption 6 providing the dominating function, and the almost sure convergence of $(\gamma_n^o, \hat{\tau}_n, \hat{\lambda}_n)$ to $(\gamma^*, \tau^*, \lambda^*)$ it follows that $\ell im_{n\to\infty}[\mathcal{I}_n(\hat{\lambda}_n), \mathcal{J}_n(\hat{\lambda}_n)] = (\mathcal{I}^* + u^*, \mathcal{J}^*)$ almost surely (Problem 7). Similar arguments apply to $\mathcal{I}_n^o$, and $\mathcal{J}_n^o$. □

As illustrated by Example 1, the usual consequence of a correctly specified model and a sensible estimation procedure is:

$$\gamma_n^{\circ} = \gamma^* \text{ for all n implies } \lambda_n^{\circ} = \lambda^* \text{ for all n .}$$

If $\lambda_n^{\circ} = \lambda^*$ for all n then we have

$$\sqrt{n}(\hat{\lambda}_n - \lambda^*) \xrightarrow{\mathcal{L}} N[\,0, (\mathcal{J}^*)^{-1}\, \mathcal{I}^* (\mathcal{J}^*)^{-1}] \text{ .}$$

But, in general, even if $\gamma_n^{\circ} \equiv \gamma^*$ for all n it is not true that

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

for some finite Δ . To reach the conclusion that $\sqrt{n}(\hat{\lambda}_n - \lambda^*)$ is asymptotically normally distributed, one must append additional regularity conditions. There are three options.

The first is to impose a Pitman drift. Estimation methods of the usual sort are designed with some class of models in mind. The idea is to imbed this intended class in a larger class $Y(e,x,\gamma)$ so that any member of the intended class is given by $Y(e,x,\gamma^*)$ for some choice of γ^* . For this choice one has

$$\gamma_n^{\circ} \equiv \gamma^* \text{ for all n implies } \lambda_n^{\circ} = \lambda^* \text{ for all n .}$$

A misspecified model would correspond to some $\gamma^{\#}$ such that $Y(e,x,\gamma^{\#})$ is outside the intended class of models. Starting with $\gamma_1^{\circ} = \gamma^{\#}$ one chooses a sequence $\gamma_2^{\circ}, \gamma_3^{\circ}, \ldots$ that converges to γ^* fast enough that $\ell\text{im}_{n\to\infty} \sqrt{n}(\lambda_n^{\circ} - \lambda^*) = \Delta$ for some finite Δ ; the most natural choice would seem to be $\Delta = 0$. See Problem 14 for the details of this construction. Since γ can be infinite dimensional, one has considerable lattitude in the choice of $Y(e,x,\gamma^{\#})$.

The second is to hold $\gamma_n^\circ \equiv \gamma^*$ and speed up the rate of convergence of Cesaro sums. If the sequence $\{x_t\}_{t=1}^\infty$ is chosen such that

$$\ell im_{n\to\infty}\sqrt{n}\,[(\partial/\partial\lambda)s_n^\circ(\lambda) - (\partial/\partial\lambda)s^*(\lambda)] = K(\lambda) \quad \text{uniformly on } \Lambda^*$$

where $K(\lambda)$ is some finite valued function then (Problem 15)

$$\ell im_{n\to\infty}\sqrt{n}\,(\lambda_n^\circ - \lambda^*) = \Delta$$

for some finite Δ. For example, if the sequence $\{x_t\}_{t=1}^\infty$ consists of replicates of T points --- that is, one puts $x_t = a_{t \bmod T}$ for some set of points $a_0, a_1, \ldots, a_{T-1}$ --- then for $i = 1, 2, \ldots, p$

$$\sup_{\Lambda^*}\sqrt{n}\,|(\partial/\partial\lambda_i)s_n^\circ(\lambda) - (\partial/\partial\lambda_i)s^*(\lambda)|$$
$$\le (\sqrt{n}/n)\sup_{\Lambda^*}\Sigma_{j=0}^{T-2}|\int_{\mathcal{E}}(\partial/\partial\lambda_i)s[Y(e,a_j,\gamma^*),a_j,\tau_n^\circ,\lambda]\,dP(e)|$$

whence $K(\lambda) \equiv 0$.

The third is to hold $\gamma_n^\circ \equiv \gamma^*$ for all n and assume that the x_t are random variables. This has the effect of imposing $\lambda_n^\circ \equiv \lambda^*$ for all n. See Section 7 for details.

Next we establish some ancillary facts regarding the constrained estimator for use in Section 5 under the assumption of a Pitman drift. Due to the Pitman drift, these results are not to be taken as an adequate theory of constrained estimation. See Section 8 for that.

ASSUMPTION 7. (Pitman drift) The sequence $\{\gamma_n^\circ\}$ is chosen such that $\lim_{n\to\infty}\sqrt{n}(\lambda_n^\circ - \lambda_n^*) = \Delta$. Moreover, $h(\lambda^*) = 0$.

THEOREM 6. Let Assumptions 1 through 7 hold. Then:

$\tilde{\lambda}_n$ converges almost surely to λ^* ,

λ_n^* converges to λ^* ,

$\mathfrak{I}$ converges almost surely to $\mathscr{I}^* + \mathscr{U}^*$,

$\mathscr{I}_n^*$ converges to $\mathscr{I}^*$,

$\tilde{\mathscr{J}}$ converges almost surely to $\mathscr{J}^*$,

$\mathscr{J}_n^*$ converges to $\mathscr{J}^*$,

$\sqrt{n}(\partial/\partial\lambda)s_n(\lambda_n^*) - \sqrt{n}(\partial/\partial\lambda)s_n^\circ(\lambda_n^*) \xrightarrow{\mathscr{L}} N(0,\mathscr{I}^*)$,

$\sqrt{n}(\partial/\partial\lambda)s_n^\circ(\lambda_n^*)$ converges to $-\mathscr{J}^*\Delta$.

PROOF. The proof that $\tilde{\lambda}_n$ converges almost surely to λ^* is nearly word for word the same as the proof of Theorem 3. The critical inequality

$$\ell im_{m\to\infty} s_{n_m}(\tilde{\lambda}_{n_m}) \leq \ell im_{m\to\infty} s_{n_m}(\lambda^*)$$

obtains by realizing that both $h(\tilde{\lambda}_{n_m}) = 0$ and $h(\lambda^*) = 0$ under the Pitman drift assumption.

The convergence properties of $\mathfrak{I}$, $\mathscr{I}_n^*$, $\tilde{\mathscr{J}}$, $\mathscr{J}_n^*$ follow directly from the convergence of $\tilde{\lambda}_n$ and λ_n^* using the argument of the proof of Theorem 5.

Since domination implies that (Problem 11)

$$(\partial/\partial\lambda)s_n^\circ(\lambda_n^*) = (1/n)\Sigma_{t=1}^n \int_{\mathcal{E}} (\partial/\partial\lambda)s[Y(e,x_t,\gamma_n^\circ),x_t,\tau_n^\circ,\lambda_n^*] \, dP(e) \; .$$

We have from Theorem 2 that

$$\sqrt{n}(\partial/\partial\lambda)s_n(\lambda_n^*) - \sqrt{n}(\partial/\partial\lambda)s_n^\circ(\lambda_n^*) \xrightarrow{\mathcal{L}} N(0,\mathcal{J}^*) \;.$$

Note that convergence of $\{\lambda_n^*\}$ to λ^* is all that is needed here; the rate $\lim_{n\to\infty}\sqrt{n}(\lambda_n^\circ - \lambda_n^*)$ is not required up to this point in the proof.

'By Taylor's theorem, recalling that $(\partial/\partial\lambda)s_n^\circ(\lambda_n^\circ) = o(n^{-\frac{1}{2}})$,

$$\sqrt{n}(\partial/\partial\lambda)s_n^\circ(\lambda_n^*) = o(1) + \bar{\bar{\mathcal{J}}}\,\sqrt{n}(\lambda^* - \lambda_n^\circ)$$

where $\bar{\bar{\mathcal{J}}}$ is similar to $\bar{\mathcal{J}}$ in the proof of Theorem 5 and converges to $\mathcal{J}^*$ for similar reasons. Since $\sqrt{n}(\lambda_n^* - \lambda_n^\circ)$ converges to $-\Delta$ by Assumption 7, the last result follows. ▯

3-3-49

PROBLEMS

1. Prove that $s_n^\circ(\lambda)$ converges uniformly to $s^*(\lambda)$ on Λ^*.

2. Hold an $\{e_t\}$ fixed for which $\ell im_{n\to\infty} \sup_{\Lambda^*} \|g_n(\lambda) - g^*(\lambda)\| = 0$ and $\ell im_{n\to\infty} \hat{\lambda}_n = \lambda^*$. Show that $\ell im_{n\to\infty} g_n(\hat{\lambda}_n) = g^*(\lambda^*)$ if $g^*(\lambda)$ is continuous.

3. Prove that λ_n° converges to λ^*.

4. Prove Part i of Lemma 2.

5. Let $(\partial/\partial\lambda')h(\lambda)$ be a matrix of order $q \times p$ with $q < p$ such that each element of $(\partial/\partial\lambda')h(\lambda)$ is continuous on an open set Λ° containing λ^*. Let $(\partial/\partial\lambda')h(\lambda)$ have rank q at $\lambda = \lambda^*$. Prove that there is an open set containing λ^* such that rank $[(\partial/\partial\lambda')h(\lambda)] = q$ for every $\lambda \in \Theta$. Hint. There is a matrix K' of order $(p - q) \times p$ and of rank $p - q$ such that

$$A(\lambda) = \begin{bmatrix} (\partial/\partial\lambda')h(\lambda) \\ K' \end{bmatrix}$$

has rank $A(\lambda^*) = p$, why? Also, $\det A(\lambda)$ is continuous and $\Theta = \{\lambda : |\det A(\lambda)| > 0\}$ is the requisite set, why?

6. Verify the claim of the first line of the proof of Theorem 5. The essence of the argument is that one could prove Theorem 5 for a set of random variables $\bar{\lambda}_n$, $\bar{\bar{\lambda}}_n$, and so on given by Lemma 2 and then $\sqrt{n}\,\hat{\lambda}_n = \sqrt{n}\,\bar{\lambda}_n + o_s(1)$, $\sqrt{n}(\partial/\partial\lambda)s_n(\lambda_n^\circ) = \sqrt{n}(\partial/\partial\lambda)s_n(\bar{\bar{\lambda}}_n) + o_s(1)$, and so on. Make this argument rigorous.

7. Use Theorem 1 to prove that $[\mathcal{I}_n(\lambda), \mathcal{J}_n(\lambda)]$ converges almost surely, uniformly on Λ and compute the uniform limit. Why does $(\gamma_n^{\circ}, \hat{\tau}_n, \hat{\lambda}_n)$ converge almost surely to $(\gamma^*, \tau^*, \lambda^*)$? Show that $[\mathcal{I}_n(\hat{\lambda}_n), \mathcal{J}_n(\hat{\lambda}_n)]$ converges almost surely to $(\mathcal{I}^*, \mathcal{J}^*)$.

8. Show that Assumption 6 suffices to dominate the elements of

$$\int_{\mathcal{E}} (\partial/\partial\lambda)s[Y(e,x,\gamma),x,\tau,\lambda]\ dP(e) \int_{\mathcal{E}} (\partial/\partial\lambda')s[Y(e,x,\gamma),x,\tau,\lambda]\ dP(e)$$

by b(x). Then apply Theorem 1 to show that $\hat{u}_n$ converges to u^*.

9. Show that if $\rho(u) = \ell n \cosh (u/2)$ and $P(e)$ is symmetric, has finite first moment, and assigns positive probability to every nonempty, open interval then $\varphi(\delta) = \int_{\mathcal{E}} \rho(e+\delta)\, dP(e)$ exists and has a unique minimum at $\delta = 0$. Hint, rewrite $\rho(u)$ in terms of exponentials and show that $\rho(u) \leq \frac{1}{2}|u|$. Use the Mean Value Theorem and the Dominated Convergence Theorem to show that $\varphi'(\delta) = \int_{\mathcal{E}} \Psi(e+\delta)\, dP(e)$. Then show that $\varphi'(0) = 0$, $\varphi'(\delta) < 0$ if $\delta < 0$, and $\varphi'(\delta) > 0$ if $\delta > 0$.

10. Suppose that $\hat{\lambda}_n$ is computed by minimizing

$$s_n(\lambda) = (1/n)\Sigma_{t=1}^{n}\rho\{[y_t - f(x_t,\lambda)]/\tau^*\}$$

where $\tau^* > 0$ is known but that the data are actually generated according to

$$y_t = g(x_t,\gamma_n^o) + e_t \ .$$

Assuming that $s_n^o(\lambda)$ has a unique minimum λ_n^o which converges to some point λ^*, compute $\mathcal{U}_n^o$, $\mathcal{J}_n^o$, and $\mathcal{I}_n^o$.

11. Prove that under Assumptions 1 through 6,

$$\int_{\mathcal{E}} (\partial/\partial\lambda) s[Y(e,x,\gamma),x,\tau,\lambda]\, dP(e)$$

$$= (\partial/\partial\lambda) \int_{\mathcal{E}} s[Y(e,x,\gamma),x,\tau,\lambda]\, dP(e) \ .$$

Hint. See the proof of Lemma 3.

12. Suppose that G_n is a matrix with $(\partial/\partial\lambda')h(\lambda_n^*)G_n = 0$ and $\ell\text{im}_{n\to\infty} G_n = G$. Show that under Assumptions 1 through 6

$$\sqrt{n}\, G_n'(\partial/\partial\lambda)s_n(\lambda_n^*) \xrightarrow{\mathcal{L}} N(0, G'\mathscr{J}^*G) \;;$$

Assumption 7 is not needed. Hint: There are Lagrange multipliers θ_n such that $(\partial/\partial\lambda')[s_n(\lambda_n^*) + \theta_n' h(\lambda_n^*)] = 0$.

13. Suppose that there is a function $\varphi(\lambda)$ such that

$$\tau = h(\lambda)$$
$$\rho = \varphi(\lambda)$$

is a once continuously differentiable mapping with a once continuously differentiable inverse

$$\lambda = \Psi(\tau,\rho) \ .$$

Put

$$g(\rho) = \Psi(0,\rho) \ ,$$
$$\rho_n^* = \varphi(\lambda_n^*) \ ,$$
$$H_n = (\partial/\partial\lambda')h(\lambda_n^*) \ ,$$
$$G_n = (\partial/\partial\rho') \ g(\rho_n^*) \ .$$

Show that G_n is the matrix required in Problem 12. Show also that $\operatorname{rank}\begin{pmatrix} G_n' \\ H_n \end{pmatrix} = p$.

14. (Construction of a Pitman drift) Fill in the missing steps and supply the necessary regularity conditions. Let

$$\lambda_n^\circ(\gamma) \text{ minimize } s_n^\circ(\gamma,\lambda) = (1/n)\Sigma_{t=1}^n \int_{\mathcal{E}} s[Y(e,x_t,\gamma),x_t,\tau_n^\circ,\lambda] \, dP(e) \ ,$$

and let

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

Suppose that there is a point γ^* in Γ such that

$$\gamma_n^\circ \equiv \gamma^* \text{ all n implies } \lambda_n^\circ(\gamma^*) = \lambda^*(\gamma^*) \text{ all n .}$$

Suppose also that Γ is a linear space and that $(\partial/\partial\alpha) \, Y(e,x,\gamma^* + \alpha\gamma^\#)$ exists for $0 \leq \alpha \leq 1$ and for some point $\gamma^\#$ in Γ. Note that Γ can be an infinite dimensional space; a directional derivative of this sort on a normed, linear space is called a Gateau derivative (Luenberger, 1969, Sec. 7.2 or Wouk, 1979, Sec. 12.1). Let

$$\gamma(\alpha) = \gamma^* + \alpha\gamma^\# \ ,$$

$$\lambda_n^\circ(\alpha) = \lambda_n^\circ[\gamma^* + \alpha\gamma^\#]$$

and

$$\lambda^*(\alpha) = \lambda^*[\gamma^* + \alpha\gamma^\#] \ .$$

Under appropriate regularity conditions, $(\partial/\partial\alpha)\lambda_n^\circ(\alpha)$ exists and can be computed from

$$0 = (1/n)\Sigma_{t=1}^n \int_{\mathcal{E}} (\partial^2/\partial\lambda\partial y')s\{Y[e,x_t,\gamma(\alpha)],x_t,\tau_n^\circ,\lambda_n^\circ(\alpha)\}(\partial/\partial\alpha)Y[e,x_t,\gamma(\alpha)] \, dP(e)$$

$$+ \ (\partial^2/\partial\lambda\partial\lambda')s_n^\circ[\gamma(\alpha), \ \lambda_n^\circ(\alpha)] \ (\partial/\partial\alpha)\lambda(\alpha) \ .$$

Again under appropriate regularity conditions,

$$\ell im_{n\to\infty} \sup_{0\leq\alpha\leq 1} \|(\partial/\partial\alpha)\lambda_n^\circ(\alpha) - (\partial/\partial\alpha)\lambda^*(\alpha)\| = 0 .$$

Then by Taylor's theorem, for i = 1, 2, ..., p ,

$$\sqrt{n}\,[\lambda_{in}^\circ(\alpha) - \lambda_{in}^\circ(0)] = \sqrt{n}\,\alpha\,(\partial/\partial\alpha)\,\lambda_{in}^\circ(\bar{\alpha}_i)$$

where $0 \leq \bar{\alpha}_i \leq \alpha$. Let $\{\alpha_n\}_{n=1}^\infty$ be any sequence such that $\ell im_{n\to\infty}\sqrt{n}\,\alpha_n = \delta$ with δ finite. Since $\lambda_n^\circ(0) = \lambda^*(0)$ for all n and $(\partial/\partial\alpha)\lambda_n^\circ(\alpha)$ converges uniformly to $(\partial/\partial\alpha)\lambda^*(\alpha)$ we have

$$\ell im_{n\to\infty}\sqrt{n}\,[\lambda_n^\circ(\alpha_n) - \lambda^*(0)] = \delta(\partial/\partial\alpha)\lambda^*(0) .$$

If the parameters of the data generating model are set to $\gamma_n^\circ = \gamma^* + \alpha_n\gamma^\#$, then

$$\ell im_{n\to\infty}\sqrt{n}\,(\lambda_n^\circ - \lambda^*) = \Delta$$

for some finite Δ as required. Note that α_n can be chosen so that $\Delta = 0$.

Suppose that the parametric constraint $h(\lambda) = 0$ can be equivalently represented as a functional dependency $\lambda = g(\rho)$; see Problem 13 or Section 6 for the construction. What is required of $g(\rho)$ so that $\ell im_{n\to\infty}\sqrt{n}\,(\rho_n^\circ - \rho^*) = \beta$? Put $\lambda_n^* = g(\rho_n^\circ)$. What is required of $g(\rho)$ so that $\ell im_{n\to\infty}\sqrt{n}\,(\lambda_n^* - \lambda^*) = \Delta^*$? Note that $\ell im_{n\to\infty}\sqrt{n}\,(\lambda_n^\circ - \lambda_n^*) = \Delta - \Delta^*$ in this case.

15. Use Taylor's theorem twice to write

$$\sqrt{n}\,[(\partial/\partial\lambda)s^*(\lambda_n^\circ) - (\partial/\partial\lambda)s_n^\circ(\lambda^*)] = [2\,\mathcal{J}^* + o(1)]\,\sqrt{n}\,(\lambda_n^\circ - \lambda^*)\ ;$$

recall that $(\partial/\partial\lambda)s^*(\lambda^*) = (\partial/\partial\lambda)s_n^\circ(\lambda_n^\circ) = 0$. Referring to the comments following Theorem 5, verify that speeding up the rate at which Cesaro sums converge will cause $\sqrt{n}\,(\hat{\lambda}_n - \lambda^*)$ to be asymptotically normally distributed.

4. METHOD OF MOMENTS ESTIMATORS

Recall that a method of moments estimator $\hat{\lambda}_n$ is defined as the solution of the optimization problem

$$\text{Minimize:} \quad s_n(\lambda) = d[m_n(\lambda), \hat{\tau}_n]$$

where $d[m,\tau]$ is a measure of the distance of m from zero, $\hat{\tau}_n$ is an estimator of nuisance parameters, and

$$m_n(\lambda) = (1/n)\Sigma_{t=1}^{n} m(y_t, x_t, \hat{\tau}_n, \lambda) \ .$$

The constrained method of moments estimator $\tilde{\lambda}_n$ is the solution of the optimization problem

$$\text{Minimize:} \quad s_n(\lambda) \quad \text{subject to } h(\lambda) = 0 \ .$$

The objective of this section is to find the almost sure limit and the asymptotic distribution of the unconstrained estimator $\hat{\lambda}_n$ under regularity conditions that do not rule out specification error. Some ancillary facts regarding the asymptotic distribution of the constrained estimator $\tilde{\lambda}_n$ under a Pitman drift are also derived for use in the later sections on hypothesis testing. This section differs from the previous section in detail but the general pattern is much the same. Accordingly the comments on motivations, regularity conditions, and results will be abbreviated. These results are due to Burguete (1980) in the main with some refinements made here to isolate the Pitman drift assumption.

As before, an example --- a correctly specified scale invariant M-estimator --- is carried throughout the discussion to illustrate how the regularity conditions may be satisfied in applications.

EXAMPLE 2. (Scale Invariant M-estimator) The data generating model is

$$y_t = f(x_t, \gamma_n^\circ) + e_t \qquad t = 1, 2, \ldots, n \,.$$

Proposal 2 of Huber (1964) leads to the moment equations

$$m_n(\lambda) = (1/n)\Sigma_{t=1}^n \begin{pmatrix} \Psi\{[y_t - f(x_t,\theta)]/\sigma\}(\partial/\partial\theta)f(x_t,\theta) \\ \Psi^2\{[y_t - f(x_t,\theta)]/\sigma\} - \beta \end{pmatrix}$$

with $\lambda = (\theta', \sigma)'$. For specificity let

$$\Psi(u) = \tfrac{1}{2} \tanh\,(u/2) \,,$$

a bounded odd function with bounded even derivative and let

$$\beta = \int \Psi^2(e)\, d\Phi(e)$$

where Φ is the standard normal distribution function. There is no preliminary estimator $\hat{\tau}_n$ with this example so the argument τ of $m(y,x,\tau,\lambda)$ is suppressed to obtain

$$m(y,x,\lambda) = \begin{pmatrix} \Psi\{[y - f(x,\theta)]/\sigma\}(\partial/\partial\theta)f(x,\theta) \\ \Psi^2\{[y - f(x,\theta)]/\sigma\} - \beta \end{pmatrix} \,.$$

The distance function is

$$d(m) = \tfrac{1}{2} m'm \,,$$

again suppressing the argument τ , whence the estimator $\hat{\lambda}_n$ is defined as that value of λ which minimizes

$$s_n(\lambda) = \tfrac{1}{2}\, m_n'(\lambda)\, m_n(\lambda) \,.$$

The error distribution P(e) is symmetric and puts positive probability on every open interval of the real line. ▯

We call the reader's attention to some heavily used notation and then state the identification condition.

NOTATION 5.

$$m_n(\lambda) = (1/n)\Sigma_{t=1}^{n} m(y_t, x_t, \hat{\tau}_n, \lambda)$$

$$m_n^\circ(\lambda) = (1/n)\Sigma_{t=1}^{n} \int_{\mathcal{E}} m[Y(e, x_t, \gamma_n^\circ), x_t, \tau_n^\circ, \lambda]\ dP(e)$$

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

$$s_n(\lambda) = d[m_n(\lambda), \hat{\tau}_n]$$

$$s_n^\circ(\lambda) = d[m_n^\circ(\lambda), \tau_n^\circ]$$

$$s^*(\lambda) = d[m^*(\lambda), \tau^*]$$

$\hat{\lambda}_n$ minimizes $s_n(\lambda)$

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

λ_n° minimizes $s_n^\circ(\lambda)$

λ_n^* minimizes $s_n^\circ(\lambda)$ subject to $h(\lambda) = 0$

λ^* minimizes $s^*(\lambda)$

ASSUMPTION 8. (Identification) The parameter γ° is indexed by n and the sequence $\{\gamma_n^{\circ}\}$ converges to a point γ^*. The sequence of nuisance parameter estimators is centered at a point τ_n° in the sense that $\sqrt{n}(\hat{\tau}_n - \tau_n^{\circ})$ is bounded in probability; the sequence $\{\tau_n^{\circ}\}$ converges to a point τ^* and $\{\hat{\tau}_n\}$ converges almost surely to τ^*. Either the solution λ^* of the equations $m^*(\lambda) = 0$ is unique or there is one solution λ^* that can be regarded as being naturally associated to γ^*. Further, $(\partial/\partial\lambda')m^*(\lambda^*)$ has full column rank $(= p)$.

The assumption that $m^*(\lambda^*) = 0$ is somewhat implausible in those misspecified situations where the range of $m_n(\lambda)$ is in a higher dimension than the domain. As sensible estimation procedures will have $m^*(\lambda^*) = 0$ if $Y(e,x,\gamma^*)$ falls into the class of models for which it was designed one could have both $m^*(\lambda^*) = 0$ and misspecification with a Pitman drift. Problem 14 of Section 3 spells out the details; see also Problems 2 and 3 of Section 2. But this is not really satisfactory. One would rather have the freedom to hold $\gamma_n^\circ \equiv \gamma^*$ for all n at some point γ^* for which $m^*(\lambda^*) \neq 0$. Such a theory is not beyond reach but it is more complicated than for the case $m^*(\lambda^*) = 0$. As we have no need of the case $m^*(\lambda^*) \neq 0$ in the sequel, we shall spare the reader these complications in the text; the more general result is given in Problem 6.

For the example, $m^*(\lambda^*) = 0$:

EXAMPLE 2. (Continued) Let σ^* solve $\int_{\mathcal{E}} \Psi^2(e/\sigma)\, dP(e) = \beta$; a solution exists since $G(\sigma) = \int_{\mathcal{E}} \Psi^2(e/\sigma)\, dP(e)$ is a continuous, decreasing function with $G(0) = 1$ and $G(\infty) = 0$. Consider putting $\lambda = (\gamma', \sigma^*)'$. With this choice

$$\int_{\mathcal{E}} m[Y(e,x,\gamma),x,(\gamma',\sigma^*)]\, dP(e)$$

$$= \int_{\mathcal{E}} m[e + f(x,\gamma),x,(\gamma',\sigma^*)]\, dP(e)$$

$$= \begin{pmatrix} \int_{\mathcal{E}} \Psi(e/\sigma^*)\, dP(e)\, (\partial/\partial\theta) f(x,\gamma) \\ \int_{\mathcal{E}} \Psi^2(e/\sigma^*)\, dP(e) - \beta \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ 0 \end{pmatrix} .$$

As the integral is zero for every x it follows that $m^*(\lambda^*) = 0$ at $\lambda^* = (\gamma^{*\prime}, \sigma^*)'$. Similarly $m_n^\circ(\lambda_n^\circ) = 0$ at $\lambda_n^\circ = (\gamma_n^{\circ\prime}, \sigma^*)$. □

The following notation defines the parameters of the asymptotic distribution of $\hat{\lambda}_n$. The notation is not as formidable as it looks; it merely consists of breaking a computation down into its component parts.

NOTATION 6.

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

$$\bar{\bar{S}}(\lambda) = \int_{\mathcal{X}}\int_{\mathcal{E}} m[Y(e,x,\gamma^*),x,\tau^*,\lambda] m' [Y(e,x,\gamma^*),x,\tau^*,\lambda]\ dP(e)\ d\mu(x) - \bar{\bar{K}}(\lambda)$$

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

$$\bar{\bar{D}}(\lambda) = (\partial^2/\partial m\partial m')\mathscr{d}[m^*(\lambda),\tau^*]$$

$$\bar{\bar{\mathscr{I}}}(\lambda) = \bar{\bar{M}}'(\lambda)\ \bar{\bar{D}}(\lambda)\ \bar{\bar{S}}(\lambda)\ \bar{\bar{D}}(\lambda)\ \bar{\bar{M}}(\lambda)$$

$$\bar{\bar{\mathscr{J}}}(\lambda) = \bar{\bar{M}}'(\lambda)\ \bar{\bar{D}}(\lambda)\ \bar{\bar{M}}(\lambda)$$

$$\bar{\bar{u}}(\lambda) = \bar{\bar{M}}'(\lambda)\ \bar{\bar{D}}(\lambda)\ \bar{\bar{K}}(\lambda)\ \bar{\bar{D}}(\lambda)\ \bar{\bar{M}}(\lambda)$$

$$\mathscr{I}^* = \bar{\bar{\mathscr{I}}}(\lambda^*),\ \mathscr{J}^* = \bar{\bar{\mathscr{J}}}(\lambda^*),\ u^* = \bar{\bar{u}}(\lambda^*)$$

$$S^* = \bar{\bar{S}}(\lambda^*),\ M^* = \bar{\bar{M}}(\lambda^*),\ D^* = \bar{\bar{D}}(\lambda^*),\ K^* = \bar{\bar{K}}(\lambda^*)$$

We illustrate the computations with the example:

EXAMPLE 2. (Continued) For $\lambda = (\gamma',\sigma^*)'$ we have

$$\int_{\mathcal{E}} m[Y(e,x,\gamma),x,\lambda]\,dP(e)\Big|_{\lambda=(\gamma',\sigma^*)'}$$

$$= \begin{pmatrix} \int_{\mathcal{E}} \Psi(e/\sigma^*)\,dP(e)\,(\partial/\partial\theta)f(x,\gamma) \\ \int_{\mathcal{E}} \Psi^2(e/\sigma^*)\,dP(e) - \beta \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

thus $\bar{\bar{K}}(\lambda^*) = 0$ whence $u^* = 0$. Further computation yields

$$\bar{\bar{S}}(\lambda^*) = \begin{pmatrix} \int_{\mathcal{E}} \Psi^2(e/\sigma^*)\,dP(e)\,\mathfrak{F}'\mathfrak{F} & 0 \\ 0' & \int_{\mathcal{E}} [\Psi^2(e/\sigma^*) - \beta]^2\,dP(e) \end{pmatrix}$$

$$\bar{\bar{M}}(\lambda^*) = \begin{pmatrix} (1/\sigma^*)\int_{\mathcal{E}} \Psi'(e/\sigma^*)\,dP(e)\,\mathfrak{F}'\mathfrak{F} & 0 \\ 0' & -2(1/\sigma^*)^2\int_{\mathcal{E}} \Psi(e/\sigma^*)\Psi'(e/\sigma^*)e\,dP(e) \end{pmatrix}$$

$$\bar{\bar{D}}(\lambda^*) = I$$

where

$$\mathfrak{F}'\mathfrak{F} = \int_{\mathcal{X}} [(\partial/\partial\theta)f(x,\theta)][(\partial/\partial\theta)f(x,\theta)]'\,d\mu(x)\Big|_{\theta=\gamma^*} .$$

As will be seen later, it is only $V^* = (\mathcal{J}^*)^{-1}\,\mathcal{I}^*(\mathcal{J}^*)^{-1}$ that is needed. Observing that $\bar{M}(\lambda)$ is invertible we have

$$
\begin{aligned}
V^* &= (\mathcal{J}^*)^{-1}\,\mathcal{I}^*(\mathcal{J}^*)^{-1} \\
&= [\bar{\bar{M}}(\lambda^*)]^{-1}\,[\bar{\bar{S}}(\lambda^*)] \quad [\bar{\bar{M}}'(\lambda^*)]^{-1} \\
&= \begin{pmatrix} (\sigma^*)^2 \dfrac{\mathcal{E}\Psi^2(e/\sigma^*)}{[\mathcal{E}\Psi'(e/\sigma^*)]^2} \; (\mathfrak{X}'\mathfrak{X})^{-1} & 0 \\ 0' & \dfrac{(\sigma^*)^4\,\mathcal{E}[\Psi^2(e/\sigma^*) - \beta]^2}{4[\mathcal{E}\,e\,\Psi(e/\sigma^*)\Psi'(e/\sigma^*)]^2} \end{pmatrix} . \quad \square
\end{aligned}
$$

In Section 5, the distributions of test statistics are characterized in terms of the following quantities:

NOTATION 7.

$$\bar{K}_n(\lambda) = (1/n)\Sigma_{t=1}^{n}\int_{\mathcal{E}} m[Y(e,x_t,\gamma_n^\circ),x_t,\tau_n^\circ,\lambda]dP(e)\int_{\mathcal{E}} m'[Y(e,x_t,\gamma_n^\circ),x_t,\tau_n^\circ,\lambda]dP(e)$$

$$\bar{S}_n(\lambda) = (1/n)\Sigma_{t=1}^{n}\int_{\mathcal{E}} m[Y(e,x_t,\gamma_n^\circ),x_t,\tau_n^\circ,\lambda]m'[Y(e,x_t,\gamma_n^\circ),x_t,\tau_n^\circ,\lambda]dP(e) - \bar{K}_n(\lambda)$$

$$\bar{M}_n(\lambda) = (1/n)\Sigma_{t=1}^{n}\int_{\mathcal{E}} (\partial/\partial\lambda')m[Y(e,x_t,\gamma_n^\circ),x_t,\tau_n^\circ,\lambda]dP(e)$$

$$\bar{D}_n(\lambda) = (\partial^2/\partial m\partial m')\ d[m_n^\circ(\lambda),\tau_n^\circ]$$

$$\bar{\mathcal{I}}_n(\lambda) = \bar{M}_n'(\lambda)\ \bar{D}_n(\lambda)\ \bar{S}_n(\lambda)\ \bar{D}_n(\lambda)\ \bar{M}_n(\lambda)$$

$$\bar{\mathcal{J}}_n(\lambda) = \bar{M}_n'(\lambda)\ \bar{D}_n(\lambda)\ \bar{M}_n(\lambda)$$

$$\bar{\mathcal{U}}_n(\lambda) = \bar{M}_n'(\lambda)\ \bar{D}_n(\lambda)\ \bar{K}_n(\lambda)\ \bar{D}_n(\lambda)\ \bar{M}_n(\lambda)$$

$$\mathcal{I}_n^\circ = \bar{\mathcal{I}}_n(\lambda_n^\circ),\ \mathcal{J}_n^\circ = \bar{\mathcal{J}}_n(\lambda_n^\circ),\ \mathcal{U}_n^\circ = \bar{\mathcal{U}}_n(\lambda_n^\circ)$$

$$\mathcal{I}_n^* = \bar{\mathcal{I}}_n(\lambda_n^*),\ \mathcal{J}_n^* = \bar{\mathcal{J}}_n(\lambda_n^*),\ \mathcal{U}_n^* = \bar{\mathcal{U}}_n(\lambda_n^*)$$

We illustrate the computations with the example:

EXAMPLE 2. (Continued) Computations similar to those for $\mathcal{I}^*$ and $\mathcal{J}^*$ yield

$$V_n^\circ = (\mathcal{J}_n^\circ)^{-1} (\mathcal{I}_n^\circ) (\mathcal{J}_n^\circ)^{-1}$$

$$= \begin{pmatrix} \dfrac{(\sigma^*)^2 \, \mathcal{E}\Psi^2(e/\sigma^*)}{[\mathcal{E}\Psi'(e/\sigma^*)]^2} \, (F'F)^{-1} & 0 \\ 0' & \dfrac{(\sigma^*)^4 \, \mathcal{E}[\Psi^2(e/\sigma^*) - \beta]^2}{4[\mathcal{E}\, e\,\Psi(e/\sigma^*)\Psi'(e/\sigma^*)]^2} \end{pmatrix}$$

where

$$F'F = (1/n)\Sigma_{t=1}^{n}[(\partial/\partial\theta)f(x_t,\theta_n^\circ)][(\partial/\partial\theta)f(x_t,\theta_n^\circ)]' \ . \quad \square$$

Some plausible estimators of $\mathcal{I}^*$ and $\mathcal{J}^*$ are as follows:

NOTATION 8.

$$S_n(\lambda) = (1/n)\Sigma_{t=1}^{n} m(y_t,x_t,\hat{\tau}_n,\lambda)m'(y_t,x_t,\hat{\tau}_n,\lambda)$$

$$M_n(\lambda) = (1/n)\Sigma_{t=1}^{n}(\partial/\partial\lambda')m(y_t,x_t,\hat{\tau}_n,\lambda)$$

$$D_n(\lambda) = (\partial^2/\partial m\partial m')d[m_n(\lambda),\hat{\tau}_n]$$

$$\mathscr{I}_n(\lambda) = M_n'(\lambda)D_n(\lambda)S_n(\lambda)D_n(\lambda)M_n(\lambda)$$

$$\mathscr{J}_n(\lambda) = M_n'(\lambda)D_n(\lambda)M_n(\lambda)$$

$$\hat{\mathscr{I}} = \mathscr{I}_n(\hat{\lambda}_n),\ \tilde{\mathscr{I}} = \mathscr{I}_n(\tilde{\lambda}_n)$$

$$\hat{\mathscr{J}} = \mathscr{J}_n(\hat{\lambda}_n),\ \tilde{\mathscr{J}} = \mathscr{J}_n(\tilde{\lambda}_n)$$

For Example 2, there are alternative choices:

EXAMPLE 2. (Continued) Reasoning by analogy with the forms that obtain from Notation 6, most would probably substitute the following estimators for those given by Notation 8:

$$\hat{S}_n = \begin{pmatrix} (1/n)\Sigma_{t=1}^{n}\Psi^2(\hat{e}_t/\hat{\sigma})\ \hat{F}'\hat{F} & 0 \\ 0' & (1/n)\Sigma_{t=1}^{n}[\Psi^2(\hat{e}_t/\hat{\sigma}) - \beta]^2 \end{pmatrix}$$

$$\hat{M}_n = \begin{pmatrix} (1/\hat{\sigma}^2)(1/n)\Sigma_{t=1}^{n}\Psi'(\hat{e}_t/\hat{\sigma})\ \hat{F}'\hat{F} & 0 \\ 0' & -(2/\hat{\sigma}^2)(1/n)\Sigma_{t=1}^{n}\Psi(e_t/\hat{\sigma})\Psi'(e_t/\hat{\sigma})\hat{e}_t \end{pmatrix}$$

$$\hat{D}_n = I$$

where

$$\hat{e}_t = y_t - f(x_t,\hat{\theta}_n)$$

$$\hat{F}'\hat{F} = (1/n)\Sigma_{t=1}^{n}[(\partial/\partial\theta)f(x_t,\hat{\theta}_n)][(\partial/\partial\theta)f(x_t,\hat{\theta}_n)]' . \quad \square$$

We shall adjoin some technical assumptions. As before, one may assume that $\hat{\tau}_n$ takes its values in a compact ball T for which τ^* is an interior point without loss of generality. Similarly for the parameter space Γ . This leaves domination as the essential condition. We have commented previously (Section 2) on the implications of a compact estimation space Λ^* . In the previous section we commented on the construction of the requisite dominating function b(e,x) .

ASSUMPTION 9. There are closed balls Λ^* and T centered at λ^* and τ^* respectively with finite, nonzero radii for which the elements of $m(y,x,\tau,\lambda),(\partial/\partial\lambda_i)m(y,x,\tau,\lambda),(\partial^2/\partial\lambda_i\partial\lambda_j)m(y,x,\tau,\lambda)$ are continuous and dominated by $b[q(y,x,\gamma),x]$ on $\mathcal{Y} \times \mathcal{X} \times T \times \Lambda^* \times \Gamma$; $b(e,x)$ is that of Assumption 3. The distance function $d(m,\tau)$ and derivatives $(\partial/\partial m)d(m,\tau)$, $(\partial^2/\partial m\partial\tau')d(m,\tau)$, $(\partial^2/\partial m\partial m')d(m,\tau)$ are continuous on $\mathfrak{J} \times T$ where $\mathfrak{J}$ is some closed ball centered at the zero vector with finite, nonzero radius.

The only distance functions that we shall ever consider have the form

$$d(m,\tau) = m'\Psi(\tau)m$$

with $\Psi(\tau)$ positive definite over T . There seems to be no reason to abstract beyond the essential properties of distance functions of this form so we impose:

ASSUMPTION 10. The distance function satisfies: $(\partial/\partial m)\, d(0,\tau) = 0$ for all τ in T (which implies $(\partial^2/\partial m \partial \tau')d(0,\tau) = 0$ for all τ in T), and $(\partial^2/\partial m \partial m')d(0,\tau)$ is positive definite for all τ in T .

If the point λ^* that satisfies $m^*(\lambda) = 0$ is unique over Λ^*, then $s^*(\lambda)$ will have a unique minimum over Λ^* for any distance function that increases with $\|m\|$. In this case the same argument used to prove Theorem 2 can be used to conclude that $\hat{\lambda}_n$ converges almost surely to λ^*. But in many applications, the moment equations are the first order conditions of an optimization problem. In these applications it is unreasonable to expect $m^*(\lambda)$ to have a unique root over some natural estimation space Λ^*. To illustrate, consider posing Example 1 as a method of moments problem:

EXAMPLE 1. (Continued) The optimization problem

$$\text{Minimize:}\quad s_n(\lambda) = (1/n)\Sigma_{t=1}^{n}\rho\,\{[y_t - f(x_t,\lambda)]/\hat{\tau}_n\}$$

has first order conditions $m_n(\hat{\lambda}_n) = 0$ with

$$m_n(\lambda) = (1/n)\Sigma_{t=1}^{n}\Psi\{[y_t - f(x_t,\lambda)]/\hat{\tau}_n\}(\partial/\partial\lambda)f(x,\lambda)\ .$$

We have seen that it is quite reasonable to expect that the almost sure limit $s^*(\lambda)$ of $s_n(\lambda)$ will have a unique minimum λ^* minimum over Λ^* . But, depending on the choice of $f(x,\theta)$, $s^*(\lambda)$ can have local minima and saddle points over Λ^* as well. In this case $m^*(\lambda)$ will have a root at λ^* but $m^*(\lambda)$ will also have roots at each local minimum and each saddle point. Thus, if Example 1 is recast as the problem

$$\text{Minimize:}\quad s_n(\lambda) = (1/2)m_n'(\lambda)m_n(\lambda)$$

we cannot reasonably assume that $s^*(\lambda)$ will have a unique minimum. □

Without the assumption that $m^*(\lambda)$ has a unique root, the best consistency result that we can obtain is that $s_n(\lambda)$ will eventually have a local minimum near λ^*. We collect together a list of facts needed throughout this section as a lemma then prove the result:

LEMMA 4. Under Assumptions 1 through 3 and 8 through 10, interchange of differentiation and integration is permitted in these instances:

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

$$(\partial^2/\partial\lambda_i\partial\lambda_j)m_\alpha^*(\lambda) = \int_{\mathcal{X}}\int_{\mathcal{E}}(\partial^2/\partial\lambda_i\partial\lambda_j)m_\alpha[Y(e,x,\gamma^*),x,\tau^*,\lambda]dP(e)\ d\mu(x)$$

$$(\partial/\partial\lambda_i)m_{\alpha n}^\circ(\lambda) = (1/n)\Sigma_{t=1}^n\int_{\mathcal{E}}(\partial/\partial\lambda_i)m_\alpha[Y(e,x,\gamma_n^\circ),x,\tau_n^\circ,\lambda]dP(e)$$

$$(\partial^2/\partial\lambda_i\partial\lambda_j)m_{\alpha n}^\circ(\lambda) = (1/n)\Sigma_{t=1}^n\int_{\mathcal{E}}(\partial^2/\partial\lambda_i\partial\lambda_j)m_\alpha[Y(e,x,\gamma_n^\circ),x,\tau_n^\circ,\lambda]dP(e)$$

There is a closed ball Λ centered at λ^* with finite nonzero radius such that:

$\ell im_{n\to\infty}m_n^\circ(\lambda) = m^*(\lambda)$ uniformly on Λ,

$\ell im_{n\to\infty}(\partial/\partial\lambda_i)m_n^\circ(\lambda) = (\partial/\partial\lambda_i)m^*(\lambda)$ uniformly on Λ,

$\ell im_{n\to\infty}(\partial^2/\partial\lambda_i\partial\lambda_j)m_n^\circ(\lambda) = (\partial^2/\partial\lambda_i\partial\lambda_j)m^*(\lambda)$ uniformly on Λ,

$\ell im_{n\to\infty}m_n(\lambda) = m^*(\lambda)$ almost surely, uniformly on Λ,

$\ell im_{n\to\infty}(\partial/\partial\lambda_i)m_n(\lambda) = (\partial/\partial\lambda_i)m^*(\lambda)$ almost surely, uniformly on Λ,

$\ell im_{n\to\infty}(\partial^2/\partial\lambda_i\partial\lambda_j)m_n(\lambda) = (\partial^2/\partial\lambda_i\partial\lambda_j)m^*(\lambda)$ almost surely, uniformly on Λ,

$\ell im_{n\to\infty}s_n^\circ(\lambda) = s^*(\lambda)$ uniformly on Λ,

$\ell im_{n\to\infty}(\partial/\partial\lambda)s_n^\circ(\lambda) = (\partial/\partial\lambda)s^*(\lambda)$ uniformly on Λ,

$\ell im_{n\to\infty}(\partial^2/\partial\lambda\partial\lambda')s_n^\circ(\lambda) = (\partial^2/\partial\lambda\partial\lambda')s^*(\lambda)$ uniformly on Λ,

$\ell im_{n\to\infty}s_n(\lambda) = s^*(\lambda)$ almost surely, uniformly on Λ,

$\ell im_{n\to\infty}(\partial/\partial\lambda)s_n(\lambda) = (\partial/\partial\lambda)s^*(\lambda)$ almost surely, uniformly on Λ,

$\ell im_{n\to\infty}(\partial^2/\partial\lambda\partial\lambda')s_n(\lambda) = (\partial^2/\partial\lambda\partial\lambda')s^*(\lambda)$ almost surely, uniformly on Λ,

and

$M^* = (\partial/\partial\lambda')m^*(\lambda^*)$,

$(\partial/\partial\lambda)s^*(\lambda^*) = 0$,

$(\partial^2/\partial\lambda\partial\lambda')s^*(\lambda^*) = \mathcal{J}^*$.

PROOF. The arguments used in the proof of Lemma 3 may be repeated to show that interchange of differentiation and integration is permitted on Λ^* and that the sequences involving $m_n^\circ(\lambda)$ and $m_n(\lambda)$ converge uniformly on Λ^*. So let us turn our attention to $s_n(\lambda) = d[m_n(\lambda), \hat{\tau}_n]$.

Differentiating, we have for $m_n(\lambda) \in \mathfrak{F}$ that

$$(\partial/\partial\lambda_i)s_n(\lambda) = \Sigma_\alpha(\partial/\partial m_\alpha)d[m_n(\lambda), \hat{\tau}_n](\partial/\partial\lambda_i)m_{\alpha n}(\lambda),$$

and

$$(\partial^2/\partial\lambda_i\partial\lambda_j)s_n(\lambda) = \Sigma_\alpha\Sigma_\beta(\partial^2/\partial m_\alpha\partial m_\beta)d[m_n(\lambda), \hat{\tau}_n](\partial/\partial\lambda_i)m_{\alpha n}(\lambda)(\partial/\partial\lambda_j)m_{\beta n}(\lambda)$$

$$+ \Sigma_\alpha(\partial/\partial m_\alpha)d[m_n(\lambda), \hat{\tau}_n](\partial^2/\partial\lambda_i\partial\lambda_j)m_{\alpha n}(\lambda) .$$

Fix a sequence $\{e_t\}$ for which $\hat{\tau}_n$ converges to τ^* and for which $m_n(\lambda)$ converges uniformly to $m^*(\lambda)$ on Λ^*; almost every $\{e_t\}$ is such. Now $m^*(\lambda^*) = 0$ by assumption and $m^*(\lambda)$ is continuous on the compact set Λ^* as it is the uniform limit of continuous functions. Thus there is a $\delta > 0$ such that

$$\|\lambda - \lambda^*\| \leq \delta \Rightarrow \|m^*(\lambda)\| < \eta$$

where η is the radius of the closed ball $\mathfrak{F}$ given by Assumption 9. Then there is an N such that

$$n > N, \ \|\lambda - \lambda^*\| \leq \delta \Rightarrow \ \|m_n(\lambda)\| < \eta .$$

Set $\Lambda = \{\lambda : \|\lambda - \lambda^*\| \leq \delta\}$.

Now $(\partial/\partial m_\alpha)d(m,\tau)$ is a continuous function on the compact set $\mathfrak{F} \times T$ so it is uniformly continuous on $\mathfrak{F} \times T$, see Problem 1. Then since $m_n(\lambda)$ converges uniformly to $m^*(\lambda)$ and $\hat{\tau}_n$ converges to τ^* it follows that $(\partial/\partial m_\alpha)d[m_n(\lambda), \hat{\tau}_n]$ converges uniformly to $(\partial/\partial m_\alpha)d[m^*(\lambda), \tau^*]$; similarly for $d[m_n(\lambda), \hat{\tau}_n]$ and $(\partial^2/\partial m_\alpha\partial m_\beta)d[m_n(\lambda), \hat{\tau}_n]$. The uniform convergence of $s_n(\lambda)$, $(\partial/\partial\lambda_i)s_n(\lambda)$ and $(\partial^2/\partial\lambda_i\partial\lambda_j)s_n(\lambda)$ follows at once. Since the convergence is uniform for almost every $\{e_t\}$ it is uniform almost surely. Similar arguments apply to $s_n^\circ(\lambda)$.

By the interchange result $M^* = (\partial/\partial\lambda')m^*(\lambda^*)$. Differentiating,

$$(\partial/\partial\lambda_i)s^*(\lambda^*) = \Sigma_\alpha(\partial/\partial m_\alpha)d[m^*(\lambda^*),\tau^*](\partial/\partial\lambda_i)m_\alpha^*(\lambda^*) \ .$$

As $m^*(\lambda^*) = 0$ and $(\partial/\partial m)d(0,\tau^*) = 0$, $(\partial/\partial\lambda)s^*(\lambda^*) = 0$. Differentiating once more,

$$\begin{aligned}&(\partial^2/\partial\lambda_i\partial\lambda_j)s^*(\lambda^*)\\ &\quad = \Sigma_\alpha\Sigma_\beta(\partial^2/\partial m_\alpha\partial m_\beta)d(0,\tau^*)(\partial/\partial\lambda_i)m_\alpha^*(\lambda^*)(\partial/\partial\lambda_j)m_\beta^*(\lambda^*)\\ &\qquad + \Sigma_\alpha(\partial/\partial m_\alpha)d(0,\tau^*)(\partial^2/\partial\lambda_i\partial\lambda_j)m_\alpha^*(\lambda^*) \ .\end{aligned}$$

The second term is zero as $(\partial/\partial m)d(0,\tau^*) = 0$ whence

$$\begin{aligned}(\partial^2/\partial\lambda\partial\lambda')s^*(\lambda) &= [(\partial/\partial\lambda')m^*(\lambda^*)]'(\partial^2/\partial m\partial m')d(0,\tau^*)(\partial/\partial\lambda')m^*(\lambda^*)\\ &= (M^*)'D^*M^* = \mathcal{J}^* \ . \ \square\end{aligned}$$

THEOREM 7. (Existence of consistent local minima) Let Assumptions 1 through 3 and 8 through 10 hold. Then there is a closed ball Λ centered at λ^* with finite, nonzero radius such that the sequence $\{\hat{\lambda}_n\}$ of $\hat{\lambda}_n$ that minimize $s_n(\lambda)$ over Λ converges almost surely to λ^* and the sequence $\{\lambda_n^o\}$ of λ_n^o that minimize $s_n^o(\lambda)$ over Λ converges to λ^*.

PROOF. By Lemma 4 and by assumption, $(\partial/\partial\lambda)s^*(\lambda^*) = 0$ and $(\partial^2/\partial\lambda\partial\lambda')s^*(\lambda^*)$ is positive definite. Then there is a closed ball Λ' centered at λ^* with finite, nonzero radius on which $s^*(\lambda)$ has a unique minimum at $\lambda = \lambda^*$ (Bartle, 1964, Sec. 21). Let Λ'' be the set given by Lemma 4 and put $\Lambda = \Lambda' \cap \Lambda''$. Then $s^*(\lambda)$ has a unique minimum on Λ and both $s_n(\lambda)$ and $s_n^o(\lambda)$ converge almost surely to $s^*(\lambda)$ uniformly on Λ. The argument used to prove Theorem 3 may be repeated here word for word to obtain the conclusions of the theorem. □

The following additional regularity conditions are needed to obtain asymptotic normality. The integral condition is similar to that in Assumption 6 ; the comments following Assumption 6 apply here as well.

ASSUMPTION 11. The elements of $m(y,x,\tau,\lambda)\ m'(y,x,\tau,\lambda)$ and $(\partial/\partial\tau')$ $m(y,x,\tau,\lambda)$ are continuous and dominated by $b[q(y,x,\gamma),x]$ on $\mathcal{Y}\times\mathcal{X}\times T\times\Lambda^*\times\Gamma$; $b(e,x)$ is that of Assumption 3. The elements of $(\partial^2/\partial\tau\partial m')d(m,\tau)$ are continuous on $\mathfrak{J}\times T$.

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

Next we show that the "scores" $(\partial/\partial\lambda)s_n(\lambda_n^\circ)$ are asymptotically normally distributed. As noted earlier, we rely heavily on the assumption that $m^*(\lambda^*) = 0$. To remove it, see Problem 6.

THEOREM 8. (Asymptotic normality of the scores) Under Assumptions 1 through 3 and 8 through 11

$$\sqrt{n}\ (\partial/\partial\lambda)s_n(\lambda_n^\circ)\xrightarrow{\mathcal{L}}N(0,\mathcal{J}^*)\ .$$

$\mathcal{J}^*$ may be singular.

PROOF. By Lemma 2, we may assume without loss of generality that $\hat{\lambda}_n$ and λ_n° lie in the smallest of the closed balls given by Assumptions 9 and 11, Lemma 4, and Theorem 7 and that $(\partial/\partial\lambda)s_n(\hat{\lambda}_n) = o_s(n^{-\frac{1}{2}})$ and $(\partial/\partial\lambda)s_n^\circ(\lambda_n^\circ) = o(n^{-\frac{1}{2}})$.

A typical element of the vector $\sqrt{n}(\partial/\partial m)\,d\,[m_n(\lambda_n^\circ),\hat{\tau}_n]$ can be expanded about $[m_n^\circ(\lambda_n^\circ),\tau_n^\circ]$ to obtain

$$\begin{aligned}\sqrt{n}\ (\partial/\partial m_\alpha)\ d[m_n(\lambda_n^\circ),\hat{\tau}_n]& \\ &= \sqrt{n}\ (\partial/\partial m_\alpha)\ d[m_n^\circ(\lambda_n^\circ),\tau_n^\circ] + (\partial/\partial\tau')(\partial/\partial m_\alpha)\ d(\bar{m},\bar{\tau})\sqrt{n}(\hat{\tau}_n - \tau_n^\circ) \\ &\quad + (\partial/\partial m')(\partial/\partial m_\alpha)\ d(\bar{m},\bar{\tau})\ \sqrt{n}\,[m_n(\lambda_n^\circ) - m_n^\circ(\lambda_n^\circ)]\end{aligned}$$

where $(\bar{m},\bar{\tau})$ is on the line segment joining $[m_n(\lambda_n^\circ),\hat{\tau}_n]$ to $[m_n^\circ(\lambda_n^\circ),\tau_n^\circ]$. Thus $(\bar{m},\bar{\tau})$ converges almost surely to (m^*,τ^*) where $m^* = m^*(\lambda^*)$. Noting that $\sqrt{n}\ (\hat{\tau}_n - \tau_n^\circ)$ is bounded in probability by Assumption 8 and that

$$\sqrt{n}\,[m_n(\lambda_n^\circ) - m_n^\circ(\lambda_n^\circ)]\xrightarrow{\mathcal{L}}N(0,S^*)$$

by Theorem 2 we may write (Problem 3)

$$\sqrt{n}\ (\partial/\partial m)\ d[m_n(\lambda_n^\circ),\ \hat{\tau}_n]$$

$$= \sqrt{n}\ (\partial/\partial m)\ d[m_n^\circ(\lambda_n^\circ), \tau_n^\circ] + (\partial^2/\partial m \partial \tau')d(m^*, \tau^*)\ \sqrt{n}(\hat{\tau}_n - \tau_n^\circ)$$

$$+ (\partial^2/\partial m \partial m')d(m^*, \tau^*)\ \sqrt{n}\ [m_n(\lambda_n^\circ) - m_n^\circ(\lambda_n^\circ)] + o_p(1)\ .$$

Then

$$\sqrt{n}(\partial/\partial\lambda)s_n(\lambda_n^\circ) = \sqrt{n}(\partial/\partial\lambda)s_n(\lambda_n^\circ) + \sqrt{n}(\partial/\partial\lambda)s_n^\circ(\lambda_n^\circ) + o\ (1)$$

$$= \sqrt{n}\ M_n'(\lambda_n^\circ)(\partial/\partial m)\ d[m_n(\lambda_n^\circ), \hat{\tau}_n] + \sqrt{n}\ \bar{M}_n'(\lambda_n^\circ)(\partial/\partial m)d[m_n^\circ(\lambda_n^\circ), \tau_n^\circ] + o_s(1)$$

$$= \sqrt{n}\ [M_n(\lambda_n^\circ) - \bar{M}_n(\lambda_n^\circ)]'(\partial/\partial m)d[m_n^\circ(\lambda_n^\circ), \tau_n^\circ]$$

$$+ M_n'(\lambda_n^\circ)\ [(\partial^2/\partial m \partial \tau')d(m^*, \tau^*)]\ \sqrt{n}(\hat{\tau}_n - \tau_n^\circ)$$

$$+ M_n'(\lambda_n^\circ)[(\partial^2/\partial m \partial m')d(m^*, \tau^*)]\ \sqrt{n}\ [m_n(\lambda_n^\circ) - m_n^\circ(\lambda_n^\circ)] + o_p(1)\ .$$

Note that by Theorem 2 $\sqrt{n}\ [M_n(\lambda_n^\circ) - \bar{M}_n(\lambda_n^\circ)]$ is also bounded in probability so that we have (Problem 3) the critical equation of the proof:

$$\sqrt{n}\ (\partial/\partial\lambda)s_n(\lambda_n^\circ) = \sqrt{n}\ [M_n(\lambda_n^\circ) - \bar{M}_n(\lambda_n^\circ)]'(\partial/\partial m)d(m^*, \tau^*)$$

$$+ (M^*)'[(\partial^2/\partial m \partial \tau')d(m^*, \tau^*)]\ \sqrt{n}\ (\hat{\tau}_n - \tau_n^\circ)$$

$$+ (M^*)'\ D^*\ \sqrt{n}\ [m_n(\lambda_n^\circ) - m_n^\circ(\lambda_n^\circ)] + o_p(1)\ .$$

We assumed that $m^* = 0$ so that the first two terms on the right hand side drop out by Assumption 10. Inspecting the third term, we can conclude at once that

$$\sqrt{n}\ (\partial/\partial\lambda)s_n(\lambda_n^\circ) \xrightarrow{\mathcal{L}} N[0, (M^*)'D^*S^*D^*M^*]\ .$$

In general the first two terms must be taken into account (Problem 6). □

Asymptotic normality of the unconstrained method of moments estimator follows at once:

THEOREM 9 . Let Assumptions 1 through 3 and 8 through 11 hold. Then:

$$\sqrt{n}(\hat{\lambda}_n - \lambda_n^\circ) \xrightarrow{\mathcal{L}} N[0, (\mathcal{J}^*)^{-1}\mathcal{I}^*(\mathcal{J}^*)^{-1}] ,$$

$\hat{\mathcal{I}}$ converges almost surely to $\mathcal{I}^* + \mathcal{U}^*$,

$\mathcal{I}_n^\circ$ converges to $\mathcal{I}^*$,

$\hat{\mathcal{J}}$ converges almost surely to $\mathcal{J}^*$,

$\mathcal{J}_n^\circ$ converges to $\mathcal{J}^*$.

$\mathcal{I}^*$ may be singular.

PROOF. By Lemma 2, we may assume without loss of generality that $\hat{\lambda}_n$, and λ_n° lie in the smallest of the closed balls given by Assumptions 9 and 11, Lemma 4, and Theorem 7 and that $(\partial/\partial\lambda)s_n(\hat{\lambda}_n) = o_s(n^{-\frac{1}{2}})$, $(\partial/\partial\lambda)s_n^\circ(\lambda_n^\circ) = o(n^{-\frac{1}{2}})$.

By Taylor's theorem and arguments similar to the previous proof

$$\sqrt{n}(\partial/\partial\lambda)s_n(\lambda_n^\circ) = \sqrt{n}(\partial/\partial\lambda)s_n(\hat{\lambda}_n) + [\mathcal{J}^* + o_s(1)]\sqrt{n}(\lambda_n^\circ - \hat{\lambda}_n)$$

$$= o_s(1) + [\mathcal{J}^* + o_s(1)]\sqrt{n}(\lambda_n^\circ - \hat{\lambda}_n) .$$

Then by Slutsky's theorem (Serfling, 1980, Sec. 1.5.4 or Rao, 1973, Sec. 2c.4)

$$\sqrt{n}(\lambda_n^\circ - \hat{\lambda}_n) \xrightarrow{\mathcal{L}} N[0, (\mathcal{J}^*)^{-1}\mathcal{I}^*(\mathcal{J}^*)^{-1}] .$$

This establishes the first result.

We shall show that $\hat{\mathcal{I}}$ converges almost surely to $\mathcal{I}^* + \mathcal{U}^*$. The arguments for $\mathcal{I}_n^\circ$, $\hat{\mathcal{J}}$, and $\mathcal{J}_n^\circ$ are similar. Now $\hat{\mathcal{I}}$ is defined as

$$\hat{\mathcal{I}} = M_n'(\hat{\lambda}_n)D_n(\hat{\lambda}_n)S_n(\hat{\lambda}_n)D_n(\hat{\lambda}_n)M_n(\hat{\lambda}_n) .$$

Since the Cesaro sum

$$(1/n)\Sigma_{t=1}^{n} m[Y(e_t,x_t,\gamma),x_t,\tau,\lambda]\, m'[Y(e_t,x_t,\gamma),x_t,\tau,\lambda]$$

converges almost surely to the integral

$$\int_{\mathcal{X}}\int_{\mathcal{E}} m[Y(e,x,\gamma),x,\tau,\lambda]\; m'[Y(e,x,\gamma),x,\tau,\lambda]\; dP(e)\; d\mu(x)$$

uniformly on $\Gamma \times T \times \Lambda$ by Theorem 1 with Assumption 11 providing the dominating function and since $(\gamma_n^{o},\hat{\tau}_n,\hat{\lambda}_n)$ converges almost surely to $(\gamma^*, \tau^*, \lambda^*)$ we have that

$$\begin{aligned} &\ell im_{n\to\infty} S_n(\hat{\lambda}_n) \\ &\quad = \int_{\mathcal{X}}\int_{\mathcal{E}} m[Y(e,x,\gamma^*),x,\tau^*,\lambda^*]\; m'[Y(e,x,\gamma^*),x,\tau^*,\lambda^*]\; dP(e)\; d\mu(x) \\ &\quad = S^* + K^* \end{aligned}$$

almost surely. A similar argument shows that $M_n(\hat{\lambda}_n)$ converges almost surely to M^*. Since $(\partial^2/\partial m\partial m')d(m,\tau)$ is continuous in (m,τ) by Assumption 9 and $[m_n(\hat{\lambda}_n),\hat{\tau}_n]$ converges almost surely to $(0,\tau^*)$ by Lemma 4, Theorem 7, and Assumption 8 we have that $D_n(\hat{\lambda}_n)$ converges almost surely to D^*. Thus

$$\begin{aligned} \ell im_{n\to\infty} \hat{\mathcal{J}} &= (M^*)'D^*(S^* + K^*)\; D^*M^* \\ &= \mathcal{J}^* + \mathcal{U}^* \end{aligned}$$

almost surely. ▯

The variance formula

$$\mathcal{J}^{-1}\mathcal{I}\,\mathcal{J}^{-1} = (M'DM)^{-1}(M'DSDM)(M'DM)^{-1}$$

is the same as that which would result if the generalized least squares estimator

$$\hat{\beta} = (M'DM)^{-1}M'Dy$$

were employed for the linear model

$$y = M\beta + e,\ e \sim (0,S) \ .$$

Thus, the greatest efficiency for given moment equations results when $D^* = (S^*)^{-1}$.

A construction of τ_n° for Example 1 was promised:

EXAMPLE 1. (Continued) Assume that f, μ, and P are such that Assumptions 1 through 6 are satisfied for the preliminary estimator $\hat{\theta}_n$. Then $\hat{\theta}_n$ has a center θ_n^o such that $\sqrt{n}(\hat{\theta}_n - \theta_n^o)$ is bounded in probability and $\ell im_{n\to\infty} \theta_n^o = \gamma^*$. Let

$$m(y,x,\theta,\tau) = \Psi^2\{[y - f(x,\theta)]/\tau\} - \int \Psi^2(e)\, d\Phi(e)$$

and

$$m_n(\tau) = (1/n)\Sigma_{t=1}^n m(y_t,x_t,\hat{\theta}_n,\tau) \,.$$

The almost sure limit of $m_n(\tau)$ is

$$m^*(\tau) = \int_{\mathcal{X}}\int_{\mathcal{E}} m[Y(e,x,\gamma^*),x,\gamma^*,\tau]\, dP(e)\, d\mu(x)$$

$$= \int_{\mathcal{E}} \Psi^2(e/\tau)\, dP(e) - \int \Psi^2(e)\, d\Phi(e) \,.$$

Since $0 < \int \Psi^2(e)d\Phi(e) < 1$ and $G(\tau) = \int_{\mathcal{E}} \Psi^2(e/\tau)\, dP(e)$ is a continuous, decreasing function with $G(0) = 1$ and $G(\infty) = 0$ there is a τ^* with $m^*(\tau^*) = 0$. Assume that f, μ, and P are such that Assumptions 8 through 11 are satisfied for $s_n(\tau) = (\frac{1}{2})\, m_n^2(\tau)$. Then by Theorem 7 and 9, $\hat{\tau}_n$ has a center τ_n^o such that $\sqrt{n}(\hat{\tau}_n - \tau_n^o)$ is bounded in probability and $\ell im_{n\to\infty}\tau_n^o = \tau^*$. ▯

The argument used in the example is a fairly general approach for verifying the regularity conditions regarding nuisance parameter estimators. Typically, a nuisance parameter estimator solves an equation of the form

$$m_n(\tau) = (1/n)\Sigma_{t=1}^{n} m(y_t, x_t, \hat{\theta}_n, \tau)$$

where $\hat{\theta}_n$ minimizes an $s_n(\theta)$ that is free of nuisance parameters. As such $\hat{\theta}_n$ comes equipped with a center θ_n^o as defined in either Section 3 or 4 . Let

$$m_n^o(\tau) = (1/n)\Sigma_{t=1}^{n} \int_{\mathcal{E}} m[Y(e, x_t, \gamma_n^o), x_t, \theta_n^o, \tau] \; dP(e)$$

let $d(m) = m'm/2$, then the appropriate center

$$\tau_n^o \text{ minimizes } s_n^o(\tau) = d[m_n^o(\tau)] \text{ .}$$

Next we establish some ancillary facts regarding the constrained estimation under a Pitman drift for use in Section 5. As noted previously, these results are not to be taken as an adequate theory of constrained estimation; that is found in Section 8.

ASSUMPTION 12. (Pitman drift) The sequence $\{\gamma_n^\circ\}$ is chosen such that $\ell im_{n\to\infty}\sqrt{n}(\lambda_n^\circ - \lambda_n^*) = \Delta$. Moreover, $h(\lambda^*) = 0$.

THEOREM 10. Let Assumptions 1 through 3 and 8 through 12 hold. Then there is a closed ball Λ centered at λ^* with finite, nonzero radius such that the constrained estimator $\tilde{\lambda}_n$ converges almost surely to λ^* and λ_n^* converges to λ^*. Moreover:

$\tilde{\mathcal{I}}$ converges almost surely to $\mathcal{I}^* + \mathcal{U}^*$,

$\mathcal{I}_n^*$ converges to $\mathcal{I}^*$,

$\tilde{\mathcal{J}}$ converges almost surely to $\mathcal{J}^*$,

$\mathcal{J}_n^*$ converges to $\mathcal{J}^*$,

$\sqrt{n}(\partial/\partial\lambda)s_n(\lambda_n^*) - \sqrt{n}(\partial/\partial\lambda)s_n^\circ(\lambda_n^*) \xrightarrow{\mathcal{L}} N(0,\mathcal{I}^*)$,

$\sqrt{n}(\partial/\partial\lambda)s_n^\circ(\lambda_n^*)$ converges to $-\mathcal{J}^*\Delta$.

PROOF. The argument showing the convergence of $\{\tilde{\lambda}_n\}$ and $\{\lambda_n^*\}$ is the same as the proof of Theorem 7 with the argument modified as per the proof of Theorem 6. The argument showing the convergence of $\tilde{\mathcal{I}}$, $\mathcal{I}_n^*$, $\tilde{\mathcal{J}}$, and $\mathcal{J}_n^*$ is the same as in the proof of Theorem 9. The same argument used in the proof of Theorem 8 may be used to derive the equation

$$\begin{aligned}
&\sqrt{n}\ (\partial/\partial\lambda)s_n(\lambda_n^*) - \sqrt{n}\ (\partial/\partial\lambda)s_n^\circ(\lambda_n^*) \\
&\quad = \sqrt{n}\ [M_n(\lambda_n^*) - \bar{M}_n(\lambda_n^*)]'(\partial/\partial m)d(m^*,\tau^*) \\
&\qquad + (M^*)'[(\partial^2/\partial m\partial\tau')d(m^*,\tau^*)]\sqrt{n}\ (\hat{\tau}_n - \tau_n^\circ) \\
&\qquad + (M^*)'D^*\sqrt{n}\ [m_n(\lambda_n^*) - m_n^\circ(\lambda_n^*)] + o_p(1)\ .
\end{aligned}$$

We assumed that $m^* = 0$ so that the first two terms on the right hand side drop out. By Theorem 2,

$$\sqrt{n}\,[m_n(\lambda_n^*) - m_n^\circ(\lambda_n^*)] \xrightarrow{\mathcal{L}} N(0,S^*)$$

whence

$$\sqrt{n}\,(\partial/\partial\lambda)s_n(\lambda_n^*) - \sqrt{n}\,(\partial/\partial\lambda)s_n^\circ(\lambda_n^*) \xrightarrow{\mathcal{L}} N[0,(M^*)'D^*S^*D^*M^*]$$

and the first result follows.

The argument that $\sqrt{n}\,(\partial/\partial\lambda)s_n^\circ(\lambda_n^*)$ converges to $-\mathcal{J}^*\Delta$ is the same as in the proof of Theorem 6. □

PROBLEMS

1. A vector valued function $f(x)$ is said to be uniformly continuous on X if given $\epsilon > 0$ there is a $\delta > 0$ such that for all x, x' in X with $\|x - x'\| < \delta$ we have $\|f(x) - f(x')\| < \epsilon$. If $f(x)$ is a continuous function and X is compact then $f(x)$ is uniformly continuous on X. (Royden, 1963, Ch. 9). Let $g_n(t)$ take its values in X and let $\{g_n(t)\}$ converge uniformly to $g(t)$ on T. Show that $\{f[g_n(t)]\}$ converges uniformly to $f[g(t)]$ on T.

2. Prove that $\lim_{n\to\infty} m_n^o(\lambda) = m^*(\lambda)$ uniformly on Λ^*. Prove that $\lim_{n\to\infty} s_n^o(\lambda) = s^*(\lambda)$ uniformly on Λ^*.

3. A (vector-valued) random variable Y_n is bounded in probability if given any $\epsilon > 0$ and $\delta > 0$ there is an M and an N such that $P(\|Y_n\| > M) < \delta$ for all $n > N$. Show that if $Y_n \xrightarrow{\mathcal{L}} N(\mu, V)$ then Y_n is bounded in probability. Show that if X_n is a random matrix each element of which converges in probability to zero and Y_n is bounded in probability then $X_n Y_n$ converges in probability to the zero vector. Hint, see Rao (1973, Sec. 2c.4).

4. Prove that $\mathcal{I}_n^o$ converges to $\mathcal{I}^*$ and that $\mathscr{J}$ converges almost surely to $\mathscr{J}^*$.

5. Compute $\bar{K}_n(\lambda)$, $\bar{M}_n(\lambda)$, and $\bar{S}_n(\lambda)$ for Example 2 in the case $\lambda \neq \lambda_n^o$.

6. Let Assumptions 1 through 3 and 8 through 11 hold except that $m^*(\lambda^*) \neq 0$; also, $(\partial/\partial m)d(0,\tau)$ and $(\partial^2/\partial m \partial \lambda')d(0,\tau)$ can be nonzero. Suppose that the nuisance parameter estimator can be written as

$$\sqrt{n}\,(\hat{\tau}_n - \tau_n^\circ) = A_n\,(1/\sqrt{n})\,\Sigma_{t=1}^{n}\,f(y_t, x_t, \theta_n^\circ) + o_p(1)$$

where $\ell im_{n\to\infty}\theta_n^\circ = \theta^*$, $\ell im_{n\to\infty} A_n = A^*$ almost surely, and $f(y,x,\theta)$ satisfies the hypotheses of Theorem 2. Let $m^* = m^*(\lambda^*)$ and define:

$$Z(e,x) = \begin{pmatrix} m[Y(e,x,\gamma^*),x,\tau^*,\lambda^*] \\ \text{vec } (\partial/\partial\lambda)m'[Y(e,x,\gamma^*),x,\tau^*,\lambda^*] \\ f[Y(e,x,\gamma^*),x,\theta^*] \end{pmatrix}$$

$$\mathcal{K}^* = \int_{\mathcal{X}} \{\int_{\mathcal{E}} Z(e,x)\;dP(e)\}\{\int_{\mathcal{E}} Z(e,x)\;dP(e)\}'d\mu(x)$$

$$\mathcal{S}^* = \int_{\mathcal{X}}\int_{\mathcal{E}} Z(e,x)\;Z'(e,x)\;dP(e)\;d\mu(x) - \mathcal{K}^*$$

$$G^* = [\,(M^*)'D^* \;\vdots\; (\partial/\partial m')d(m^*,\tau^*) \otimes I_p \;\vdots\, (M^*)'(\partial^2/\partial m\partial\tau')d(m^*,\tau^*)A^*]$$

$$\mathcal{I}^* = G^*\,\mathcal{S}^*(G^*)'$$

$$\mathcal{J}^* = (\partial^2/\partial\lambda\partial\lambda')s^*(\lambda^*)$$

Show that

$$\sqrt{n}\;s_n(\lambda_n^\circ) \xrightarrow{\mathcal{L}} N(0,\mathcal{I}^*)\;,$$

$$\sqrt{n}\;(\hat{\lambda}_n - \lambda_n^\circ) \xrightarrow{\mathcal{L}} N[\,0,(\mathcal{J}^*)^{-1}\;\mathcal{I}^*(\mathcal{J}^*)^{-1}]\;.$$

Hint: Recall that if A of order r by c is partitioned as $A = [a_1 \vdots\, a_2 \vdots\, \ldots\, \vdots\, a_c]$ then

$$\text{vec } A = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_c \end{pmatrix}$$

and vec AB = $(B' \otimes I_r)$ vec A where $\otimes$ denotes the Kronecker product of two matrices. See the proofs of Theorems 8 and 9 .

7. Under the same assumptions as in Problem 6 show that

$$\sqrt{n}\ (\partial/\partial\lambda)s_n(\lambda_n^*) - \sqrt{n}\ (\partial/\partial\lambda)s_n^\circ(\lambda_n^*) \xrightarrow{\mathcal{L}} N(0,\mathcal{J}^*)$$

where $\mathcal{J}^*$ is defined as in Problem 6.

5. TESTS OF HYPOTHESES

Both paths lead to the same results. At first the path follows Assumptions 1 through 3 which describe the data generating process. Then the road forks. One can follow the least mean distance estimation path with Notations 1 through 4 defining the quantities:

$\hat{\lambda}_n$, $\tilde{\lambda}_n$, λ_n°, and λ_n^* ;

$s_n(\lambda)$ and $s_n^{\circ}(\lambda)$;

$\hat{\mathscr{I}}$, $\tilde{\mathscr{I}}$, $\mathscr{I}_n^{\circ}$, and $\mathscr{I}_n^*$

$\hat{\mathscr{J}}$, $\tilde{\mathscr{J}}$, $\mathscr{J}_n^{\circ}$, and $\mathscr{J}_n^*$

$\mathscr{U}_n^{\circ}$, and $\mathscr{U}_n^*$.

Or, one can follow the method of moments,estimation path with Notations 5 through 8 defining these quantities. In either case the results are the same and may be summarized as follows:

SUMMARY. Let Assumptions 1 through 3 hold and let either Assumptions 4 through 7 or 8 through 12 hold. Then on a closed ball Λ centered at λ^* with finite, nonzero radius:

$s_n(\lambda)$ and $s_n^\circ(\lambda)$ converge almost surely and uniformly on Λ to $s^*(\lambda)$;

$(\partial/\partial\lambda)s_n(\lambda)$ and $(\partial/\partial\lambda)s_n^\circ(\lambda)$ converge almost surely and uniformly on Λ to $(\partial/\partial\lambda)s^*(\lambda)$;

$(\partial^2/\partial\lambda\partial\lambda')s_n(\lambda)$ and $(\partial^2/\partial\lambda\partial\lambda')s_n^\circ(\lambda)$ converge almost surely and uniformly on Λ to $(\partial^2/\partial\lambda\partial\lambda')s^*(\lambda)$, and $(\partial^2/\partial\lambda\partial\lambda')s^*(\lambda^*) = \mathcal{J}^*$;

$\sqrt{n}(\partial/\partial\lambda)s_n(\lambda_n^*) - \sqrt{n}(\partial/\partial\lambda)s_n^\circ(\lambda_n^*) \xrightarrow{\mathcal{L}} N(0,\mathcal{I}^*)$;

$\sqrt{n}(\hat{\lambda}_n - \lambda_n^\circ) \xrightarrow{\mathcal{L}} N[0,(\mathcal{J}^*)^{-1}\mathcal{I}^*(\mathcal{J}^*)^{-1}]$;

$\sqrt{n}(\lambda_n^\circ - \lambda_n^*)$ converges to Δ, $\sqrt{n}(\partial/\partial\lambda)s_n^\circ(\lambda_n^*)$ converges to $-\mathcal{J}^*\Delta$;

$\hat{\lambda}_n$ and $\tilde{\lambda}_n$ converge almost surely to λ^* and $h(\lambda^*) = 0$;

$\hat{\mathcal{I}}$ and $\tilde{\mathcal{I}}$ converge almost surely to $\mathcal{I}^* + \mathcal{U}^*$, $\mathcal{I}_n^\circ$ and $\mathcal{I}_n^*$ converge to $\mathcal{I}^*$;

$\hat{\mathcal{J}}$ and $\tilde{\mathcal{J}}$ converge almost surely to $\mathcal{J}^*$, $\mathcal{J}_n^\circ$ and $\mathcal{J}_n^*$ converge to $\mathcal{J}$.

$\mathcal{U}_n^\circ$ and $\mathcal{U}_n^*$ converge to $\mathcal{U}^*$

Taking the Summary as the point of departure, consider testing

$$H\colon h(\lambda_n^\circ) = 0 \quad \text{against } A\colon h(\lambda_n^\circ) \neq 0 \,.$$

Three tests for this hypothesis will be studied: the Wald test, Rao's efficient score test (Lagrange multiplier test), and an analog of the likelihood ratio test. 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 9.

$$V_n^\circ = (\mathcal{J}_n^\circ)^{-1}\, \mathcal{I}_n^\circ (\mathcal{J}_n^\circ)^{-1},\ V_n^* = (\mathcal{J}_n^*)^{-1}\, \mathcal{I}_n^* (\mathcal{J}_n^*)^{-1}$$

$$\hat{V} = \hat{\mathcal{J}}^{-1}\, \hat{\mathcal{I}}\, \hat{\mathcal{J}}^{-1},\ \tilde{V} = \tilde{\mathcal{J}}^{-1}\, \tilde{\mathcal{I}}\, \tilde{\mathcal{J}}^{-1}$$

$$\hat{h} = h(\hat{\lambda}),\ H(\lambda) = (\partial/\partial\lambda')h(\lambda)$$

$$H_n^\circ = H(\lambda_n^\circ),\ H_n^* = H(\lambda_n^*)$$

$$\hat{H} = H(\hat{\lambda}_n),\ \tilde{H} = H(\tilde{\lambda}_n)$$

Theorem 11:

$$V = V_n^\circ,\ \mathcal{I} = \mathcal{I}_n^\circ,\ \mathcal{J} = \mathcal{J}_n^\circ,\ \mathcal{U} = \mathcal{U}_n^\circ,\ H = H_n^\circ$$

Theorems 12, 13, 14, and 15:

$$V = V_n^*,\ \mathcal{I} = \mathcal{I}_n^*,\ \mathcal{J} = \mathcal{J}_n^*,\ \mathcal{U} = \mathcal{U}_n^*,\ H = H_n^*$$

ASSUMPTION 13. The function $h(\lambda)$ that defines the null hypothesis $H\colon h(\lambda_n^o) = 0$ is a once continuously differentiable mapping of the estimation space into R^q. Its Jacobian $H(\lambda) = (\partial/\partial\lambda')h(\lambda)$ has full rank (=q) at $\lambda = \lambda^*$. The matrix $V = \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}$ has full rank. The statement "the null hypothesis is true" means that $h(\lambda_n^o) = 0$ for all n or, equivalently, that $\lambda_n^o = \lambda_n^*$ for all n sufficiently large.

The first statistic considered is the Wald test statistic

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

which is the same idea as division of an estimator by its standard error or studentization. The statistic is simple to compute and may be computed solely from the results of an unconstrained optimization of $s_n(\lambda)$. It has two disadvantages. The first, its asymptotic distribution is a poorer approximation to its small sample distribution than for the next two statistics if Monte Carlo simulations are any guide (Chapter 1). The second, it is not invariant to reparameterization. With the same data and an equivalent model and hypotheses, two investigators could obtain different values of the test statistic. (Problem 6).

The second statistic considered is Rao's efficient score test statistic

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

Since $(\partial/\partial\lambda)[s_n(\tilde{\lambda}_n) + \tilde{\theta}_n' h(\tilde{\lambda}_n)] = 0$ for large n, an alternative form is

$$R = n\,\tilde{\theta}_n'\tilde{H}\,\mathcal{J}^{-1}\,\tilde{H}'(\tilde{H}\tilde{V}\tilde{H}')^{-1}\,\tilde{H}\mathcal{J}^{-1}\,\tilde{H}'\tilde{\theta}_n\ ,$$

which gives rise to the term Lagrange multiplier test. Quite often $V = \mathcal{I}^{-1} = \mathcal{J}^{-1}$ so that $\tilde{\mathcal{I}}^{-1}$ could be substituted for $\tilde{V}$ and $\mathcal{J}^{-1}$ in these formulas resulting in a material simplification. The statistic may be computed solely from a constrained optimization of $s_n(\lambda)$. Often, the minimization of $s_n(\lambda)$ subject to $h(\lambda) = 0$ is considerably easier than an unconstrained minimization; H: $\lambda_n^{\circ} = 0$ for example. In these cases R is easier to compute than W. There are several motivations for the statistic R of which the simplest is probably the following. Suppose that the quadratic surface

$$q(\lambda) = s_n(\tilde{\lambda}_n) + (\partial/\partial\lambda')s_n(\tilde{\lambda}_n)(\lambda - \tilde{\lambda}_n) + \tfrac{1}{2}(\lambda - \tilde{\lambda}_n)'\tilde{\mathcal{J}}(\lambda - \tilde{\lambda}_n)$$

is an accurate approsimation to the surface $s_n(\lambda)$ over a region that includes $\hat{\lambda}_n$. The quadratic surface is minimized at

$$\lambda = \tilde{\lambda}_n - \mathcal{J}^{-1}(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)$$

so that

$$\tilde{\lambda}_n - \hat{\lambda}_n \doteq \mathcal{J}^{-1}(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) .$$

Thus, $\mathcal{J}^{-1}(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)$ is the difference between $\tilde{\lambda}_n$ and $\hat{\lambda}_n$ induced by the constraint $h(\lambda) = 0$ and R is a measure of the squared length of this difference. Stated differently, $\mathcal{J}^{-1}(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)$ is a full Newton iterative step from $\tilde{\lambda}_n$ (presumably) toward $\hat{\lambda}_n$ and R is a measure of the step-length.

The third test statistic considered is an analog of the likelihood ratio test

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

The statistic measures the increase in the objective function due to the constraint $h(\tilde{\lambda}_n) = 0$; one rejects for large values of L . The statistic is derived by treating $s_n(\lambda)$ as if it were the negative of the log likelihood and applying the definition of the likelihood ratio test.

Our plan is to derive approximations to the sampling distributions of these three statistics that are reasonably accurate in applications. To illustrate the ideas as we progress, we shall carry along a misspecified model as an example:

EXAMPLE 3. One fits the nonlinear model

$$y_t = f(x_t,\lambda) + u_t\,, \quad t = 1,2,\ldots, n$$

by least squares to data that actually follow the model

$$y_t = g(x_t,\gamma_n^\circ) + e_t\,, \quad t = 1,2,\ldots, n$$

where the errors e_t are independently distributed with mean zero and variance σ^2. The hypothesis of interest is

$$H: \tau_n^\circ = \tau^* \text{ against } A: \tau_n^\circ \neq \tau^*$$

where

$$\lambda = (\rho', \tau')'$$

ρ is an r-vector, and τ is a q-vector with $p = r+q$. As in Chapter 1, we can put the model in a vector form:

$$y = f(\lambda) + u,\ y = g(\gamma_n^\circ) + e\,,\ F(\lambda) = (\partial/\partial\lambda')f(\lambda)\,.$$

We shall presume throughout that this model satisfies Assumptions 1 through 7, and 13. Direct computation yields:

$$s(y_t,x_t,\lambda) = [y_t - f(x_t,\lambda)]^2\,,$$

$$s_n(\lambda) = (1/n)\,\|y - f(\lambda)\|^2 = (1/n)[y - f(\lambda)]'[y - f(\lambda)]\,,$$

$$(\partial/\partial\lambda)s_n(\lambda) = (-2/n)F'(\lambda)[y - f(\lambda)]\,,$$

$$s_n^\circ(\lambda) = \sigma^2 + (1/n)\,\|g(\gamma_n^\circ) - f(\lambda)\|^2\,,$$

$$(\partial/\partial\lambda)s_n^\circ(\lambda) = (-2/n)F'(\lambda)[g(\gamma_n^\circ) - f(\lambda)]\,,$$

λ_n° minimizes $(1/n)\,\|g(\gamma_n^\circ) - f(\lambda)\|^2$,

$\tilde{\rho}_n$ minimizes $(1/n)\,\|y - f(\rho,\tau^*)\|^2$,

$$\tilde{\lambda}_n = (\tilde{\rho}_n',\tau^{*\prime})',$$

ρ_n^* minimizes $(1/n)\,\|g(\gamma_n^\circ) - f(\rho,\tau^*)\|^2$,

$\lambda_n^* = (\rho_n^{*\prime},\tau^{*\prime})'$,

$F_n^\circ = F(\lambda_n^\circ)$, $F_n^* = F(\lambda_n^*)$, $\hat{F} = F(\hat{\lambda}_n)$, $\tilde{F} = F(\tilde{\lambda}_n)$

$\mathscr{I}(\lambda) = (4\sigma^2/n)F'(\lambda)F(\lambda)$

$\mathscr{J}(\lambda) = (2/n)F'(\lambda)F(\lambda) - (2/n)\Sigma_{t=1}^n[g(x_t,\gamma_n^\circ) - f(x_t,\lambda)](\partial^2/\partial\lambda\partial\lambda')f(x_t,\lambda)$,

$\mathscr{U}(\lambda) = (4/n)\Sigma_{t=1}^n[g(x_t,\gamma_n^\circ) - f(x_t,\lambda)]^2[(\partial/\partial\lambda)f(x_t,\lambda)][(\partial/\partial\lambda)f(x_t,\lambda)]'$,

$\hat{\mathscr{I}} = (4/n)\Sigma_{t=1}^n[y_t - f(x_t,\hat{\lambda}_n)]^2[(\partial/\partial\lambda)f(x_t,\hat{\lambda}_n)][(\partial/\partial\lambda)f(x_t,\hat{\lambda}_n)]'$,

$\hat{\mathscr{J}} = (2/n)\hat{F}'\hat{F} - (2/n)\Sigma_{t=1}^n[y_t - f(x_t,\hat{\lambda}_n)](\partial^2/\partial\lambda\partial\lambda')f(x_t,\hat{\lambda}_n)$,

$H = [0 \vdots I_q]$, I_q is the identity matrix of order q .

The estimator

$$\hat{V} = (\hat{\mathscr{J}})^{-1}\,\hat{\mathscr{I}}(\hat{\mathscr{J}})^{-1}$$

obtained according to the general theory is not that customarily used in nonlinear regression analysis as we have seen in Chapter 1. It has an interesting property in that if the model is correctly specified, that is γ and λ have the same dimension and $g(x,\gamma) = f(x,\gamma)$, then $\hat{V}$ will yield the correct standard errors for $\hat{\lambda}_n$ even if $\mathrm{Var}(e_t) = \sigma^2(x_t)$. White (1980) terms $\hat{V}$ the heteroscadastic invariant estimator of the variance-covariance matrix of $\hat{\lambda}_n$ for this reason.

The estimator customarily employed is

$$\hat{\Omega} = n\,s^2(\hat{F}'\hat{F})^{-1}$$

with

$$s^2 = (n-p)^{-1}\|y - f(\hat{\lambda}_n)\|^2 .$$

We shall substitute $\hat{\Omega}$ for $\hat{V}$ in what follows mainly to illustrate how the general theory is to be modified to accomodate special situations. □

The limiting distributions that have been derived thus far have been stated in terms of the parameters $\mathcal{I}^*$, $\mathcal{J}^*$, and $\mathcal{U}^*$. To use these results, it is necessary to compute $\mathcal{I}^*$, $\mathcal{J}^*$, and $\mathcal{U}^*$ and to compute them it is necessary to specify the limit of $\hat{\lambda}_n$ and $\hat{\tau}_n$ and to specify the limiting measure μ on $\mathcal{X}$. Most would prefer to avoid the arbitrariness resulting from having to specify what is effectively unknowable in any finite sample. More appealing is to center $\hat{\lambda}_n$ at λ_n^o rather than at λ^*, center $\hat{\tau}_n$ at τ_n^o , and use the empirical distribution function computed from $\{x_t\}_{t=1}^n$ to approximate μ . What results is $\mathcal{I}_n^o$, $\mathcal{J}_n^o$, and $\mathcal{U}_n^o$ as approximations to $\mathcal{I}^*$, $\mathcal{J}^*$, and $\mathcal{U}^*$. The next theorem uses a Skorokhod representation to lend some formality to this approach in approximating the finite sample distribution of W. For the example we need an approximation to the limit of $\hat{\Omega}$:

EXAMPLE 3 (Continued)

The almost sure limit of $\hat{\Omega}$ is

$$\Omega^* = \{\sigma^2 + \int_{\mathcal{X}} [g(x,\gamma^*) - f(x,\lambda^*)]^2 d\mu(x)\}[\int_{\mathcal{X}} (\partial/\partial\lambda) f(x,\lambda^*) d\mu(x)]^{-1} .$$

Following the same logic that leads to the approximation of $\mathcal{S}^*$, $\mathcal{J}^*$ and $\mathcal{U}^*$ by $\mathcal{S}_n^\circ$, $\mathcal{J}_n^\circ$, and $\mathcal{U}_n^\circ$ we obtain

$$\Omega(\lambda) = n[\sigma^2 + \frac{1}{n} \|g(\gamma_n^\circ) - f(\lambda)\|^2][F'(\lambda)F(\lambda)]^{-1}$$

and

$$\Omega_n^\circ = n(\sigma^2 + \frac{1}{n} \|g(\gamma_n^\circ) - f(\lambda_n^\circ)\|^2)(F_n^{\circ\prime} F_n^\circ)^{-1}$$

where $F_n^\circ = F(\lambda_n^\circ)$. $\square$

THEOREM 11. Let Assumptions 1 through 3 hold and let either Assumptions 4 through 7 or 8 through 12 hold. Let

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

Under Assumption 10,

$$W \sim Y + o_p(1)$$

where

$$Y = Z'[H\mathcal{J}^{-1}(\mathcal{I}+\mathcal{U})\mathcal{J}^{-1}H']^{-1}Z$$

and

$$Z \sim N[\sqrt{n}\ h(\lambda_n^\circ) \ , \ HVH'] \ .$$

Recall: $V = V_n^\circ$, $\mathcal{I} = \mathcal{I}_n^\circ$, $\mathcal{J} = \mathcal{J}_n^\circ$, $\mathcal{U} = \mathcal{U}_n^\circ$, and $H = H_n^\circ$. If $\mathcal{U} = 0$ then Y has the non-central chi-square distribution with q degrees of freedom and non-centrality parameter $\alpha = nh'(\lambda_n^\circ)(HVH')^{-1}h(\lambda_n^\circ)/2$. Under the null hypothesis $\alpha = 0$.

PROOF. By Lemma 2, we may assume without loss of generality that $\hat{\lambda}_n, \lambda_n^\circ \in \Lambda$ and that $(\partial/\partial\lambda)s_n(\hat{\lambda}_n) = o_s(n^{-\frac{1}{2}}), (\partial/\partial\lambda)s_n^\circ(\lambda_n^\circ) = o(n^{\frac{1}{2}})$. By Taylor's theorem

$$\sqrt{n}\ [h_i(\hat{\lambda}_n) - h_i(\lambda_n^\circ)] = (\partial/\partial\lambda')h_i(\bar{\lambda}_{in})\sqrt{n}(\hat{\lambda}_n - \lambda_n^\circ) \qquad i = 1,2,\ldots,q$$

where $\|\bar{\lambda}_{in} - \lambda_n^\circ\| \leq \|\hat{\lambda}_n - \lambda_n^\circ\|$. By the almost sure convergence of λ_n° and $\hat{\lambda}_n$ to λ^*, $\ell im_{n\to\infty}\|\bar{\lambda}_{in} - \lambda^*\| = 0$ almost surely whence $\ell im_{n\to\infty}(\partial/\partial\lambda)h_i(\bar{\lambda}_{in}) = (\partial/\partial\lambda)h_i(\lambda^*)$ almost surely. Thus we may write

$$\sqrt{n}\ [h(\hat{\lambda}_n) - h(\lambda_n^\circ)] = [H^* + o_s(1)]\sqrt{n}(\hat{\lambda}_n - \lambda_n^\circ) \ .$$

Since $\sqrt{n}(\hat{\lambda}_n - \lambda_n^\circ) \xrightarrow{\mathcal{L}} N(0,V^*)$, we have

$$\sqrt{n}\ [h(\hat{\lambda}_n) - h(\lambda_n^\circ)] \xrightarrow{\mathcal{L}} N(0,H^*V^*H^{*\prime}) \ .$$

By Problem 3, $\ell im_{n\to\infty}\sqrt{n}\, h(\lambda_n^\circ) = H\Delta$ so that $\sqrt{n}\, h(\hat{\lambda}_n)$ is bounded in probability. Now $\hat{H}\hat{V}\hat{H}'$ converges almost surely to $H^*(\mathcal{J}^*)^{-1}(\mathcal{I}^*+\mathcal{U}^*)(\mathcal{J}^*)^{-1}H^{*\prime}$ which is nonsingular whence

$$(\hat{H}\hat{V}\hat{H}')^{-1} = [H\mathcal{J}^{-1}(\mathcal{I}+\mathcal{U})\mathcal{J}^{-1}H']^{-1} + o_s(1) \ .$$

Then

$$W = n\ h'(\hat{\lambda}_n)[H\mathcal{J}^{-1}(\mathcal{I}+\mathcal{U})\mathcal{J}^{-1}H']^{-1}h(\hat{\lambda}_n) + o_p(1) \ .$$

By the Skorokhod representation theorem (Serfling, 1980, Sec. 1.6), there are random variables Y_n with the same distribution as $\sqrt{n}\, h(\hat{\lambda}_n)$ such that $Y_n - \sqrt{n}\, h(\lambda_n^\circ) = Y + o_s(1)$ where $Y \sim N(0, H^*V^*H^{*\prime})$. Factor $H^*V^*H^*$ as $H^*V^*H^{*\prime} = P^*P^{*\prime}$ and for large n factor $HVH' = QQ'$ (Problem 1). Then

$$Y_n = \sqrt{n}\, h(\lambda_n^\circ) + Q(P^*)^{-1}\, Y + [I - Q(P^*)^{-1}]Y + o_s(1) \ .$$

Since Y is bounded in probability and $[I - Q(P^*)^{-1}] = o_s(1)$ (Problem 1) we have

$$Y_n = \sqrt{n}\, h(\lambda_n^\circ) + Q(P^*)^{-1}\, Y + o_p(1)$$

where $Q(P^*)^{-1}\, Y \sim N(0, HVH')$. Let $Z = \sqrt{n}\, h(\lambda_n^\circ) + Q(P^*)^{-1}\, Y$ and the result follows. ▯

Occasionally in the literature one sees an alternative form of the Wald test statistic

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

The alternative form is obtained from the approximation $\hat{h} \doteq \hat{H}(\hat{\lambda}_n - \tilde{\lambda}_n)$ which is derived as follows. By Taylor's theorem

$$h(\hat{\lambda}_n) = h(\tilde{\lambda}_n) + \bar{H}(\hat{\lambda}_n - \tilde{\lambda}_n)$$

where $\bar{H}$ has rows $(\partial/\partial\lambda')h_i(\lambda)$ and $\bar{\lambda}$ is one the line segment joining $\tilde{\lambda}_n$ to $\hat{\lambda}_n$. By noting that $h(\tilde{\lambda}_n) = 0$ and approximating $\bar{H}$ by $\hat{H}$ one has that $\hat{h} \doteq \hat{H}(\hat{\lambda}_n - \tilde{\lambda}_n)$. Any solution of $h(\lambda) = 0$ with $(\partial/\partial\lambda')H(\lambda) \doteq \hat{H}$ would serve as well as $\tilde{\lambda}_n$ by this logic and one sees other choices at times.

As seen from Theorem 11, an asymptoticly level α test in a correctly specified situation is to reject H: $h(\lambda_n^\circ) = 0$ when W exceeds the upper $\alpha \times 100\%$ critical point of a chi-square random variable with q degrees of freedom. In a conditional analysis of an incorrectly specified situation, μ , $h(\lambda_n^\circ)$, and α will usually be non zero so nothing can be said in general. One has a quadratic form in normally distributed random variables. Direct computation for a specified $q(y,x,\gamma_n^\circ)$ is required, see Section   for details. We illustrate with the example.

EXAMPLE 3 (Continued) The hypothesis of interest is

$$H\colon \tau_n^\circ = \tau^* \text{ against } A\colon \tau_n^\circ \neq \tau^*$$

where

$$\lambda = (\rho', \tau')' \ .$$

Substituting $\hat{\Omega}$ for $\hat{V}$ the Wald statistic is

$$W = (\hat{\tau}_n - \tau^*)'[H(\hat{F}'\hat{F})^{-1}H']^{-1}(\hat{\tau}_n - \tau^*)/s^2$$

where $H = [0 \vdots I_q]$. Thus $H(\hat{F}'\hat{F})^{-1}H'$ is the submatrix of $(\hat{F}'\hat{F})^{-1}$ formed by deleting the first r rows and columns of $(\hat{F}'\hat{F})^{-1}$. W is distributed as

$$W \sim Y + o_p(1)$$

where:

$$Y = Z'[H(\tfrac{1}{n} F_n^{\circ\prime} F_n^\circ)^{-1}H']^{-1}Z/(\sigma^2 + \tfrac{1}{n}\|g(\gamma_n^\circ) - f(\lambda_n^\circ)\|^2) \ ,$$

$$Z \sim N[\sqrt{n}(\tau_n^\circ - \tau^*),\ HVH']$$

$$V = \mathscr{J}^{-1}\mathscr{I}\,\mathscr{J}^{-1}$$

$$\mathscr{I} = (4\sigma^2/n)F_n^{\circ\prime}F_n^\circ$$

$$\mathscr{J} = (2/n)F_n^{\circ\prime}F_n^\circ - (2/n)\Sigma_{t=1}^n[g(x_t,\gamma_n^\circ) - f(x_t,\lambda_n^\circ)](\partial^2/\partial\lambda\partial\lambda')f(x_t,\lambda_n^\circ) \ .$$

If the model is correctly specified then $g(x_t,\gamma_n^\circ) = f(x_t,\lambda_n^\circ)$ and these equations simplify to:

$$Y = Z'[H(\tfrac{1}{n} F_n^{\circ\prime} F_n^\circ)^{-1}H']^{-1}Z/\sigma^2 \ ,$$

$$Z \sim N[\sqrt{n}(\tau_n^\circ - \tau^*),\ \sigma^2 H(\tfrac{1}{n} F_n^{\circ\prime} F_n^\circ)^{-1}H']$$

whence Y is distributed as a non-central chi-square random variable with q degrees of freedom and non-centrality parameter

$$\alpha = (\tau_n^\circ - \tau^*)'[H(F'F)^{-1}H']^{-1}(\tau_n^\circ - \tau^*)/\sigma^2 \ . \quad \square$$

The statistic R is a quadratic form in $(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)$ and, for n large enough that $\mathcal{J}_n(\lambda)$ can be inverted in a neighborhood of $\tilde{\lambda}_n$, the statistic L is also a quadratic form in $(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)$ (Problem 8). Thus, a characterization of the distribution of $(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)$ is needed. We shall divide this derivation into two steps. First (Theorem 12), a characterization of $(\partial/\partial\lambda)s_n(\lambda_n^*)$ is obtained. Second (Theorem 13), $(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)$ is characterized as a projection of $(\partial/\partial\lambda)s_n(\lambda_n^*)$ into the column space of H_n^* .

THEOREM 12. Let Assumptions 1 through 3 hold and let either Assumptions 4 through 7 or 8 through 12 hold. Then

$$\sqrt{n}(\partial/\partial\lambda)s_n(\lambda_n^*) \sim X + o_p(1)$$

where

$$X \sim N[\sqrt{n}(\partial/\partial\lambda)s_n^\circ(\lambda_n^*), \mathcal{J}_n^*] \, .$$

PROOF. By either Theorem 6 or Theorem 10,

$$\sqrt{n}(\partial/\partial\lambda)s_n(\lambda_n^*) - \sqrt{n}(\partial/\partial\lambda)s_n^\circ(\lambda_n^*) \xrightarrow{\mathcal{L}} N(0,\mathcal{J}^*) \, .$$

By the Skorokhod representation theorem (Serfling, 1980, Sec. 1.6), there are random variables Y_n with the same distribution as $\sqrt{n}(\partial/\partial\lambda)s_n(\lambda_n^*)$ such that $Y_n - \sqrt{n}(\partial/\partial\lambda)s_n^\circ(\lambda_n^*) = Y + o_s(1)$ where $Y \sim N(0,\mathcal{J}^*)$. Then factor $\mathcal{J}^*$ as $\mathcal{J}^* = P^*(P^*)'$ and for large n factor $\mathcal{J}_n^*$ as $\mathcal{J}_n^* = QQ'$ (Problem 1) then

$$Y_n = \sqrt{n}(\partial/\partial\lambda)s_n^\circ(\lambda_n^*) + Q(P^*)^{-1}Y + [I - Q(P^*)^{-1}]Y + o_s(1) \, .$$

Let $X = \sqrt{n}(\partial/\partial\lambda)s_n^\circ(\lambda_n^*) + Q(P^*)^{-1}Y$ whence $X \sim N[\sqrt{n}(\partial/\partial\lambda)s_n^\circ(\lambda_n^*), \mathcal{J}_n^*]$ and, since $\ell im_{n\to\infty} Q = P^*$ (Problem 1), $[I - Q(P^*)^{-1}]Y = o_p(1)$. □

Note from Theorem 5 that if $s_n(\lambda)$ corresponds to a least mean distance estimator without nuisance parameters then $\mathcal{J}_n^*$ is the exact, finite sample variance of $\sqrt{n}\ (\partial/\partial\lambda)s_n(\lambda_n^*)$. In this case, Theorem 12 is no more than a suggestion that the exact variance of $\sqrt{n}\ (\partial/\partial\lambda)s_n(\lambda_n^*)$ be used in computations instead of the asymptotic variance. Next we characterize $\sqrt{n}\ (\partial/\partial\lambda)s_n(\tilde{\lambda}_n)$.

THEOREM 13. Let Assumptions 1 through 3 hold and let either Assumptions 4 through 7 or 8 through 12 hold. Under Assumption 13,

$$\sqrt{n}\;(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) = H'(H\mathcal{J}^{-1}H')^{-1}H\mathcal{J}^{-1}\sqrt{n}\;(\partial/\partial\lambda)s_n(\lambda_n^*) + o_p(1)$$

where $\mathcal{J} = \mathcal{J}_n^*$ and $H = H_n^*$.

PROOF. By Lemma 2, we may assume without loss of generality that $\hat{\lambda}_n$, $\tilde{\lambda}_n$, λ_n^o, $\lambda_n^* \in \Lambda$. By Taylor's theorem

$$\sqrt{n}\;(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) = \sqrt{n}\;(\partial/\partial\lambda)s_n(\lambda_n^*) + \bar{\mathcal{J}}\sqrt{n}\;(\tilde{\lambda}_n - \lambda_n^*)$$

$$\sqrt{n}\;h(\tilde{\lambda}_n) = \sqrt{n}\;h(\lambda_n^*) + \bar{H}\,\sqrt{n}\;(\tilde{\lambda}_n - \lambda_n^*)$$

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

$$(\partial/\partial\lambda')(\partial/\partial\lambda_i)s_n(\bar{\lambda}_{in}) \qquad i = 1,2,\ldots,p$$

and $\bar{H}$ has rows

$$(\partial/\partial\lambda')h_j(\bar{\lambda}_{jn}) \qquad j = 1,2,\ldots,q$$

with $\bar{\lambda}_{in}$ and $\bar{\lambda}_{jn}$ on the line segment joining $\tilde{\lambda}_n$ to λ_n^*. By Lemma 2, $\sqrt{n}\;h(\tilde{\lambda}_n) = o_s(1)$. Recalling that $\sqrt{n}\;h(\lambda_n^*) \equiv 0$, we have $\bar{H}\sqrt{n}\;(\tilde{\lambda}_n - \lambda_n^*) = o_s(1)$. Since $\tilde{\lambda}_n$ and λ_n^* converge almost surely to λ^*, each $\bar{\lambda}_{in}$, $\bar{\lambda}_{jn}$ converges almost surely to λ^* whence $\bar{\mathcal{J}}$ converges almost surely to $\mathcal{J}^*$ by the uniform almost sure convergence of $(\partial^2/\partial\lambda\partial\lambda')s_n(\lambda)$; $\bar{H}$ converges almost surely to H^* by the continuity of $H(\lambda)$. Thus $\bar{\mathcal{J}} = \mathcal{J} + o_s(1)$ and $\bar{H} = H + o_s(1)$. Moreover, there is an N corresponding to almost every realization of $\{e_t\}$ such that $\det(\bar{\mathcal{J}}) > 0$ for all $n > N$. Defining $\bar{\mathcal{J}}^{-1}$ arbitrarily when $\det(\bar{\mathcal{J}}) = 0$ we have

$$\mathcal{J}^{-1}\bar{\mathcal{J}}\,\sqrt{n}\;(\tilde{\lambda}_n - \lambda_n^*) \equiv \sqrt{n}\;(\tilde{\lambda}_n - \lambda_n^*)$$

for all $n > N$. Thus, $\bar{\mathcal{J}}^{-1}\bar{\mathcal{J}}\,\sqrt{n}\;(\tilde{\lambda}_n - \lambda_n^*) = \sqrt{n}\;(\tilde{\lambda}_n - \lambda_n^*) + o_s(1)$. Combining these observations, we may write

$$\bar{H}\sqrt{n}\,(\tilde{\lambda}_n - \lambda_n^*) = o_s(1)$$

$$\sqrt{n}\,(\tilde{\lambda}_n - \lambda_n^*) = \bar{\mathcal{J}}^{-1}\sqrt{n}\,(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) - \bar{\mathcal{J}}^{-1}\sqrt{n}\,(\partial/\partial\lambda)s_n(\lambda_n^*) + o_s(1)$$

whence

$$\bar{H}\bar{\mathcal{J}}^{-1}\sqrt{n}\,(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) = \bar{H}\bar{\mathcal{J}}^{-1}\sqrt{n}\,(\partial/\partial\lambda)s_n(\lambda_n^*) + o_s(1) \; .$$

Since $\sqrt{n}\,(\partial/\partial\lambda)s_n(\lambda_n^*)$ converges in distribution it is bounded in probability whence

$$\bar{H}\bar{\mathcal{J}}^{-1}\sqrt{n}\,(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) = H\mathcal{J}^{-1}\sqrt{n}\,(\partial/\partial\lambda)s_n(\lambda_n^*) + o_p(1) \; .$$

By Lemma 2, there is a sequence of Lagrange multipliers $\tilde{\theta}_n$ such that

$$\sqrt{n}\,(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) + \tilde{H}'\sqrt{n}\,\tilde{\theta}_n = o_s(1) \; .$$

Substituting into the previous equation we have

$$\bar{H}\bar{\mathcal{J}}^{-1}\tilde{H}\sqrt{n}\,\tilde{\theta}_n = H\mathcal{J}^{-1}\sqrt{n}\,(\partial/\partial\lambda)s_n(\lambda_n^*) + o_p(1) \; .$$

By Slutsky's theorem (Serfling, 1980, Sec. 1.5.4 or Rao, 1973, Sec. 2c.4), $\sqrt{n}\,\tilde{\theta}_n$ converges in distribution. In consequence, both $\sqrt{n}\,\tilde{\theta}_n$ and $\sqrt{n}\,(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)$ are bounded in probability and we have

$$\begin{aligned}
&H'(H\mathcal{J}^{-1}H')^{-1}H\mathcal{J}^{-1}\sqrt{n}\,(\partial/\partial\lambda)s_n(\lambda_n^*) \\
&\quad = H'(H\mathcal{J}^{-1}H')^{-1}H\mathcal{J}^{-1}\sqrt{n}\,(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) + o_p(1) \\
&\quad = -H'(H\mathcal{J}^{-1}H')^{-1}H\mathcal{J}^{-1}H'\sqrt{n}\,\hat{\theta}_n + o_p(1) \\
&\quad = -H'\sqrt{n}\,\tilde{\theta}_n + o_p(1) \\
&\quad = \sqrt{n}\,(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) + o_p(1) \; . \quad \square
\end{aligned}$$

A characterization of the distribution of the statistic R follows immediately from Theorem 14:

THEOREM 14. Let Assumptions 1 through 3 hold and let either Assumptions 4 through 7 or 8 through 12 hold. Let

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

Under Assumption 13

$$R \sim Y + o_p(1)$$

where

$$Y = Z'\mathcal{J}^{-1}H'[H\mathcal{J}^{-1}(\mathcal{I}+\mathcal{U})\mathcal{J}^{-1}H']^{-1}H\mathcal{J}^{-1}Z$$

and

$$Z \sim N[\sqrt{n}\ (\partial/\partial\lambda)s_n^o(\lambda_n^*),\mathcal{I}] .$$

Recall: $V = V_n^*$, $\mathcal{I} = \mathcal{I}_n^*$, $\mathcal{J} = \mathcal{J}_n^*$, $\mathcal{U} = \mathcal{U}_n^*$, and $H = H_n^*$.

If $\mathcal{U} = 0$ then Y has the non-central chi-square distribution with q degrees of freedom and non-centrality parameter $\alpha = n[(\partial/\partial\lambda)s_n^o(\lambda_n^*)]'\mathcal{J}^{-1}H'(H\,V\,H')^{-1}H\mathcal{J}^{-1} \times [(\partial/\partial\lambda)s_n^o(\lambda_N^*)]/2$. Under the null hypothesis, $\alpha = 0$.

PROOF. By Lemma 2, we may assume without loss of generality that $\tilde{\lambda}_n \in \Lambda$. By the Summary,

$$\tilde{\mathcal{J}}^{-1}\tilde{H}'(\tilde{H}\tilde{V}\tilde{H}')^{-1}\tilde{H}\tilde{\mathcal{J}}^{-1} = \mathcal{J}^{-1}H'[H\mathcal{J}^{-1}(\mathcal{I}+\mathcal{U})\mathcal{J}^{-1}H']^{-1}H\mathcal{J}^{-1} + o_s(1) .$$

By Lemma 5, $\sqrt{n}\ (\partial/\partial\lambda)s_n(\tilde{\lambda}_n)$ is bounded in probability whence we have

$$\begin{aligned} R &= n[(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)]'\mathcal{J}^{-1}H'[H\mathcal{J}^{-1}(\mathcal{I}+\mathcal{U})\mathcal{J}^{-1}H']^{-1}H\mathcal{J}^{-1}[(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)] + o_p(1) \\ &= n[(\partial/\partial\lambda)s_n(\lambda_n^*)]'\mathcal{J}^{-1}H'[H\mathcal{J}^{-1}(\mathcal{I}+\mathcal{U})\mathcal{J}^{-1}H']^{-1}H\mathcal{J}^{-1}[(\partial/\partial\lambda)s_n(\lambda_n^*)] + o_p(1) . \end{aligned}$$

The distributional result follows by Theorem 13. The matrix $\mathcal{J}^{-1}H'[H\mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}H']^{-1}H\mathcal{J}^{-1}\mathcal{I}$ is idempotent so Y follows the non-central chi-square distribution if $\mathcal{U} = 0$. □

The remarks following Theorem 11 apply here as well. In a correctly specified situation one rejects H: $h(\lambda_n^\circ) = 0$ when R exceeds the upper $\alpha \times 100\%$ critical point of a chi-square random variable with q degrees of freedom. Under correct specification and A: $h(\lambda_n^\circ) \neq 0$ then one approximates the distribution of R with the non-central chi-square distribution. Under misspecification one must approximate with a quadratic form in normally distributed random variables.

In many applications $\tilde{\mathcal{J}}^{-1} = a\tilde{V}$ for some scalar multiple a . In this event the statistic R can be put in a simpler form as follows. Since $\text{rank}(\tilde{H}) = q$ and $\tilde{H}$ is q by p one can always find a matrix $\tilde{G}$ of order p by r with $\textbf{rank}(\tilde{G}) = r = p - q$ and $\tilde{H}\tilde{G} = 0$. For such $\tilde{G}$ we shall show in the next section that

$$\tilde{H}'(\tilde{H}\tilde{V}\tilde{H}')^{-1}\tilde{H} = \tilde{V}^{-1} - \tilde{V}^{-1}\tilde{G}(\tilde{G}'\tilde{V}^{-1}\tilde{G})^{-1}\tilde{G}'\tilde{V}^{-1} .$$

Recalling that there are Lagrange multipliers $\tilde{\theta}_n$ such that

$$(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) = \tilde{H}'\tilde{\theta}_n$$

we have

$$\tilde{G}'\tilde{V}^{-1}\tilde{\mathcal{J}}^{-1}(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) = a\,\tilde{G}'\tilde{H}'\tilde{\theta}_n = 0 .$$

Consequently we may substitute $\tilde{V}^{-1}$ for $\tilde{H}'(\tilde{H}\tilde{V}\tilde{H}')^{-1}\tilde{H}$ in the formula for R to obtain the simpler form

$$R = a^2 n[(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)]'\tilde{V}[(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)] .$$

We illustrate with Example 3:

EXAMPLE 3 (Continued) Substituting

$$\tilde{\Omega} = n\,\tilde{s}^2(\tilde{F}'\tilde{F})^{-1}$$

with

$$\tilde{s}^2 = (n-p+q)^{-1}\|y - f(\tilde{\lambda}_n)\|^2$$

for $\tilde{V}$ and substituting

$$\tilde{\mathscr{J}} = (2/n)(\tilde{F}'\tilde{F})$$

for $\mathscr{J}$. We have

$$\tilde{\mathscr{J}}^{-1} = (2\tilde{s}^2)^{-1}\tilde{\Omega}$$

whence

$$R = (1/\tilde{s}^2)[y - f(\tilde{\lambda}_n)]'\tilde{F}(\tilde{F}'\tilde{F})^{-1}\tilde{F}'[y - f(\tilde{\lambda}_n)] .$$

Putting

$$F_n^* = F(\lambda_n^*) ,$$

R is distributed as

$$R \sim Y + o_p(1)$$

where

$$Y = Z'J^{-1}H'(H\Omega H')^{-1}HJ^{-1}Z$$

$$J^{-1} = (n/2)(F_n^{*\prime}F_n^*)^{-1}$$

$$\Omega = n(\sigma^2 + \frac{1}{n}\|g(\gamma_n^\circ) - f(\lambda_n^*)\|^2)(F_n^{*\prime}F_n^*)^{-1}$$

$H = [0 \vdots I_q]$, I_q is the identity matrix of order q ,

$$Z \sim N[(-2/\sqrt{n})F_n^{*\prime}[g(\gamma_n^\circ) - f(\lambda_n^*)], \mathscr{J}_n^*] ,$$

$$\mathscr{J}_n^* = (4\sigma^2/n)F_n^{*\prime}F_n^* .$$

When the model is correctly specified, these equations reduce to

$$Y \sim a_n(n/4\sigma^2)Z'(F_n^{*\prime}F_n^*)^{-1}H'[H(F_n^{*\prime}F_n^*)^{-1}H']^{-1}H(F_n^{*\prime}F_n^*)^{-1}Z$$

$$Z \sim N[(-2/\sqrt{n})\, F_n^{*\prime}[f(\lambda_n^\circ) - f(\lambda_n^*)],\ (4\sigma^2/n)(F_n^{*\prime}F_n^*)]$$

$$a_n = \sigma^2/[\sigma^2 + \tfrac{1}{n}\|\ f(\lambda_n^\circ) - f(\lambda_n^*)\|^2]\ .$$

Y/a_n is distributed as a non-central chi-square random variable with q degrees of freedom and non-centrality parameter

$$\alpha = [f(\lambda_n^\circ) - f(\lambda_n^*)]'F_n^*(F_n^{*\prime}F_n^*)^{-1}H'[H(F_n^{*\prime}F_n^*)^{-1}H']^{-1}H(F_n^{*\prime}F_n^*)^{-1}F_n^{*\prime}[f(\lambda_n^\circ) - f(\lambda_n^*)]/(2\sigma^2)$$

The non-centrality parameter may be put in the form (Problem 9)

$$\alpha = [f(\lambda_n^\circ) - f(\lambda_n^*)]'F_n^*(F_n^{*\prime}F_n^*)^{-1}F_n^{*\prime}[f(\lambda_n^\circ) - f(\lambda_n^*)]/(2\sigma^2)\ .$$

Let $\chi^{2\prime}(t|q,\alpha)$ denote the probability that a non-central chi-square random variable with q degrees of freedom and non-centrality parameter α exceeds t . One approximates the probability that R rejects H: $\tau_n^\circ = \tau^*$ at critical point c by

$$\begin{aligned} P(R > c) &\doteq P(Y > c) \\ &= P(Y/a_n > c/a_n) \\ &= \chi^{2\prime}(c/a_n|q,\alpha)\ . \end{aligned}$$

In applications, the critical point is chosen so that $\chi^{2\prime}(c|q,0) = .05$, say, and since $c/a_n > c$ when $\alpha \neq 0$ the power of the test is reduced from that which could be achieved if $a_n \equiv 1$. If

$$s^2 = (n-p)^{-1}\|y - f(\hat{\lambda}_n)\|^2$$

is substituted for $\tilde{s}^2$ in computing R then $a_n \equiv 1$. Thus, even though the computation of s^2 entails an extra minimization to obtain $\hat{\lambda}_n$, it is probably worth the bother in most instances in order to obtain the increase in power. □

THEOREM 15. Let Assumptions 1 through 3 hold and let either Assumptions 4 through 7 or 8 through 12 hold. Let

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

Under Assumption 13,

$$L \sim Y + o_p(1)$$

where

$$Y = Z'\mathcal{J}^{-1}H'(H\mathcal{J}^{-1}H')^{-1}H\mathcal{J}^{-1}Z$$

and

$$Z \sim N[\sqrt{n}\,(\partial/\partial\lambda)s_n^{\circ}(\lambda_n^*),\ \mathcal{I}] .$$

Recall: $V = V_n^*$, $\mathcal{I} = \mathcal{I}_n^*$, $\mathcal{J} = \mathcal{J}_n^*$, $\mathcal{U} = \mathcal{U}_n^*$, and $H = H_n^*$.

If $HVH' = H\mathcal{J}^{-1}H'$ then Y has the non-central chi-square distribution with q degrees of freedom and non-centrality parameter $\alpha = n(\partial/\partial\lambda)s_n^{\circ}(\lambda_n^*)\mathcal{J}^{-1}H'(H\mathcal{J}^{-1}H')^{-1}H\mathcal{J}^{-1}(\partial/\partial\lambda)s_n^{\circ}(\lambda_n^*)/2$. Under the null hypothesis, $\alpha = 0$.

PROOF. By Lemma 2 we may assume without loss of generality that $\hat{\lambda}_n, \tilde{\lambda}_n \in \Lambda$. By Taylor's theorem

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

where $\|\bar{\lambda}_n - \hat{\lambda}_n\| \le \|\tilde{\lambda}_n - \hat{\lambda}_n\|$. By the Summary, $(\tilde{\lambda}_n, \hat{\lambda}_n)$ converges almost surely to (λ^*, λ^*) and $(\partial^2/\partial\lambda\partial\lambda')s_n(\lambda)$ converges almost surely uniformly to $s^*(\lambda)$ uniformly on Λ which implies $(\partial^2/\partial\lambda\partial\lambda')s_n(\bar{\lambda}_n) = [\mathcal{J} + o_s(1)]$. By Lemma 2, $2n[(\partial/\partial\lambda)s_n(\hat{\lambda}_n)]'(\tilde{\lambda}_n - \hat{\lambda}_n) = o_s(1)$ whence

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

Again by Taylor's theorem

$$[\mathcal{J} + o_s(1)]\sqrt{n}(\tilde{\lambda}_n - \hat{\lambda}_n) = \sqrt{n}\,(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)\ .$$

Then by Slutsky's theorem (Serfling, 1980, Sec. 1.5.4 or Rao, 1973, Sec. 2c.4) $\sqrt{n}(\tilde{\lambda}_n - \hat{\lambda}_n)$ converges in distribution and is therefore bounded.

Thus

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

$$\sqrt{n}(\tilde{\lambda}_n - \hat{\lambda}_n) = \mathcal{J}^{-1}\sqrt{n}(\partial/\partial\lambda)s_n(\tilde{\lambda}_n) + o_p(1)$$

whence

$$2n[s_n(\tilde{\lambda}_n) - s_n(\hat{\lambda}_n)] = n[(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)]'\mathcal{J}^{-1}[(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)] + o_p(1)$$

and the distributional result follows at once from Theorem 13. To see that Y is distributed as the non-central chi-square when $HVH' = H\mathcal{J}^{-1}H'$ note that $\mathcal{J}^{-1}H'(H\mathcal{J}^{-1}H')^{-1}H\mathcal{J}^{-1}\mathcal{J}$ is idempotent under this condition. □

The remarks immediately following Theorems 11 and 14 apply here as well. One rejects when L exceeds the upper $\alpha \times 100\%$ critical point of a chi-square with q degrees of freedom and so on.

In the event that $\mathcal{J} = a\,\mathcal{I} + o(1)$ for some scalar multiple a, the "likelihood ratio test statistic" can be modified as follows. Let $\hat{a}_n$ be a random variable that converges either almost surely or in probability to a. Then

$$\hat{a}_n L \sim aY + o_p(1)$$

where

$$a\,Y = Z'\mathcal{I}^{-1}H'(H\mathcal{I}^{-1}H')^{-1}H\mathcal{I}^{-1}Z \;.$$

Since $\mathcal{I}^{-1}H'(H\mathcal{I}^{-1}H')H\mathcal{I}^{-1}\mathcal{I}$ is an idempotent matrix, $a\,Y$ is distributed as the non-central chi-square distribution with q degrees of freedom and non-centrality parameter

$$\alpha = n(\partial/\partial\lambda')s_n^{\circ}(\lambda_n^*)\mathcal{I}^{-1}H'(H\mathcal{I}^{-1}H')^{-1}H\mathcal{I}^{-1}(\partial/\partial\lambda)s_n^{\circ}(\lambda_n^*)/2 \;.$$

We illustrate with the example:

EXAMPLE 3 (Continued) Assuming that the model is correctly specified,

$$\mathscr{I} = (4\sigma^2/n)F_n^{*\prime}F_n^*$$

$$\mathscr{J} = (2/n)F_n^{*\prime}F_n^* - (2/n)\Sigma_{t=1}^n[f(x_t,\lambda_n^\circ) - f(x_t,\lambda_n^*)](\partial^2/\partial\lambda\partial\lambda')f(x_t,\lambda_n^*) \ .$$

By Taylor's theorem, for some $\bar{\lambda}_n$ on the line segment joining λ_n° to λ_n^*

$$\sqrt{n}\ a_{ij} = \sqrt{n}\ (2/n)\Sigma_{t=1}^n[f(x_t,\lambda_n^\circ) - f(x_t,\lambda_n^*)](\partial^2/\partial\lambda_i\partial\lambda_j)f(x_t,\lambda_n^*)$$

$$= \sqrt{n}\ (\lambda_n^\circ - \lambda_n^*)'(2/n)\Sigma_{t=1}^n(\partial/\partial\lambda)f(x_t,\bar{\lambda}_n)(\partial^2/\partial\lambda_i\partial\lambda_j)f(x_t,\lambda_n^*)$$

whence

$$\ell im_{n\to\infty}\sqrt{n}\ a_{ij} = 2\ \Delta'\int_{\mathcal{X}}(\partial/\partial\lambda)f(x,\lambda^*)(\partial^2/\partial\lambda_i\partial\lambda_j)f(x,\lambda^*)d\mu(x)$$

by Theorem 1. Thus we have

$$\mathscr{J} = (2\sigma^2)^{-1}\ \mathscr{I} + \mathcal{O}(1/\sqrt{n}) \ .$$

An estimator of σ^2 is

$$s^2 = (n - p)^{-1}\|y - f(\hat{\lambda}_n)\|^2 \ .$$

The modified "likelihood ratio test statistic" is

$$(2s^2)^{-1}L = (2s^2)^{-1}(2n)[(1/n)\|y - f(\tilde{\lambda}_n)\|^2 - (1/n)\|y - f(\hat{\lambda}_n)\|^2]$$

$$= [\|y - f(\tilde{\lambda}_n)\|^2 - \|y - f(\hat{\lambda}_n)\|^2]/s^2 \ .$$

A further division by q would convert $(2s^2)^{-1}L$ to the F-statistic of the previous chapter. Assuming correct specification, $(2s^2)^{-1}L$ is distributed to within $o_p(1)$ as the non-central chi-square distribution with q degrees of freedom and non-centrality parameter (Problem 9)

$$\alpha = [f(\lambda_n^\circ) - f(\lambda_n^*)]'F_n^*(F_n^{*\prime}F_n^*)^{-1}F_n^{*\prime}[f(\lambda_n^\circ) - f(\lambda_n^*)]/(2\sigma^2) \ .$$

Under specification error

$$(2s^2)^{-1}L \sim a\,Y + o_p(1)$$

where:

$$a\,Y = Z'\mathcal{J}^{-1}H'(H\mathcal{J}^{-1}H')^{-1}H\mathcal{J}^{-1}Z/(2\sigma^2 + \frac{2}{n}\,\|g(\gamma_n^\circ) - f(\lambda_n^\circ)\|^2)\ ,$$

$$\mathcal{J} = (2/n)F_n^{*}{}'F_n^{*} - (2/n)\Sigma_{t=1}^{n}[g(x_t,\gamma_n^\circ) - f(x_t,\lambda_n^{*})]^2(\partial^2/\partial\lambda\partial\lambda')f(x_t,\lambda_n^{*})$$

$$Z \sim N[(-2/\sqrt{n})F_n^{*}{}'[g(\gamma_n^\circ) - f(\lambda_n^{*})]\ ,\ (4\sigma^2/n)F_n^{*}{}'F_n^{*}]\ .\ \square$$

PROBLEMS

1. (Cholesky factorization) The validity of the argument in the proof of Theorems 11and13 depends on the fact that it is possible to factor a symmetric, positive definite matrix A as A = R′R in such a way that R is a continuous function of the elements of the matrix A. To see that this is so observe that

$$A = \left[\begin{array}{c|c} a_{11} & a'_{(1)} \\ \hline a_{(1)} & A_{22} \end{array}\right]$$

$$= \left[\begin{array}{c|c} r_{11} & 0 \\ \hline r_{12} & \\ \vdots & I \\ r_{1p} & \end{array}\right] \left[\begin{array}{c|c} 1 & 0 \\ \hline & \\ 0 & D_1 \\ & \end{array}\right] \left[\begin{array}{c|c} r_{11} & r_{12}\cdots r_{1p} \\ \hline & \\ 0 & I \\ & \end{array}\right]$$

where

$$\begin{aligned} r_{11} &= \sqrt{a_{11}} \\ r_{1k} &= a_{1k}/\sqrt{a_{11}} \qquad\qquad k = 2,\ldots,p \\ D_1 &= A_{22} - (1/a_{11})a_{(1)}a'_{(1)} \;. \end{aligned}$$

The r_{1k} are continuous elements of A and D_1 is a symmetric, positive definite matrix whose elements are continuous functions of the elements of A, why? This same argument can be applied to D_1 to obtain

$$A = \left[\begin{array}{cc|c} r_{11} & 0 & 0 \\ r_{12} & r_{22} & \\ \hline r_{13} & r_{23} & \\ \vdots & \vdots & I \\ r_{1p} & r_{2p} & \end{array}\right] \left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & \\ \hline & & \\ 0 & & D_2 \\ & & \end{array}\right] \left[\begin{array}{cc|c} r_{11} & r_{12} & r_{13}\cdots r_{1p} \\ 0 & r_{22} & r_{23}\cdots r_{2p} \\ \hline & & \\ 0 & & I \\ & & \end{array}\right]$$

with continuity preserved. Supply the missing steps. This argument can be repeated a finite number of times to obtain the result. The recursion formula for the Cholesky square root method is

$$r_{1k} = a_{1k}/r_{11} \qquad k = 1,2,3,\ldots,p$$

$$r_{jk} = (1/r_{jj})(a_{jk} - \Sigma_{i=1}^{j-1} r_{ij}r_{ik}) \qquad \begin{array}{l} j = 2,3,\ldots,p \\ k = j, j+1,\ldots,p \end{array}$$

Observe that on a computer A can be factored in place using only the upper triangle of A with this recursion.

2. Suppose that θ_n^o converges to θ^* and $\hat{\theta}_n$ converges almost surely to θ^*. Let $g(\theta)$ be defined on an open set Θ and let $g(\theta)$ be continuous at $\theta^* \epsilon \Theta$. Define $g(\theta)$ arbitrarily off Θ. Show that

$$g(\hat{\theta}_n) = g(\theta_n^o) + o_s(1) \quad .$$

Let θ be a square matrix and $g(\theta)$ a matrix valued function giving the inverse of θ when it exists. If θ^* is nonsingular show that there is an open neighborhood Θ about θ^* where each $\theta \epsilon \Theta$ is nonsingular and show that $g(\theta)$ is continuous at θ^*. Hint: Use $\|\theta\| = [\Sigma_{ij}\theta_{ij}^2]^{\frac{1}{2}}$, $\|g\| = [\Sigma_{ij}g_{ij}^2]^{\frac{1}{2}}$, the determinant of a matrix is continuous, and an inverse is the product of the adjoint matrix and the reciprocal of the determinant. Show that if $\sqrt{n}\,(\partial/\partial\lambda)s_n(\lambda_n^*)$ converges in distribution then $\sqrt{n}\,(\partial/\partial\lambda)s_n(\lambda_n^*)$ is bounded in probability. Show that

$$[H + o_s(1)][\mathcal{J} + o_s(1)]^{-1}\sqrt{n}\,(\partial/\partial\lambda)s_n(\lambda_n^*) = H\mathcal{J}^{-1}\sqrt{n}\,(\partial/\partial\lambda)s_n(\lambda_n^*) + o_p(1) \quad .$$

3. Expand $\sqrt{n}\,h(\lambda_n^o)$ in a Taylor's series and show that $\ell im_{n\to\infty}\sqrt{n}\,h(\lambda_n^o) = H\Delta$.

4. Verify that if the linear model

$$y_t = x_t'\beta + e_t$$

is estimated by least squares from data that actually follows

$$y_t = g(x_t, \gamma_n^o) + e_t$$

with e_t independently and normally distributed and one tests the linear hypothesis

$$H\colon R\beta = r \quad \text{against} \quad A\colon R\beta \neq r$$

then

$$\sqrt{n}\,(\partial/\partial\lambda)s_n(\lambda_n^*) \sim N[\sqrt{n}\,(\partial/\partial\lambda)s_n^o(\lambda_n^*), \mathcal{J}_n^*]\,.$$

That is, Theorem 12 gives the correct answer, not an approximation.

5. Verify that $\alpha = 0$ when the null hypothesis is true in Theorem 14.

6. (Invariance) Consider a least mean distance estimator

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

and the hypothesis

$$H\colon \lambda_n^o = \lambda^* \text{ against } A\colon \lambda_n^o \neq \lambda^*\,.$$

Let $g(\rho)$ be a twice differentiable function with twice differentiable inverse $\rho = \varphi(\lambda)$. Then an equivalent formulation of the problem is

$$\hat{\rho} \text{ minimizes } s_n[g(\rho)] = (1/n)\Sigma_{t=1}^n\, s[y_t, x_t, \hat{\tau}_n, g(\rho)]$$

and the hypothesis

$$H\colon \rho_n^o = \varphi(\lambda^*) \text{ against } A\colon \rho_n^o \neq \varphi(\lambda^*)\,.$$

Show that the computed value of the Wald test statistic can be different for these two equivalent problems. Show that the computed value of the Lagrange multiplier and "likelihood ratio" test statistics are invariant to this reparameterization.

7. (Equivalent local power) Suppose that $\mathcal{I}^* = \mathcal{J}^*$ and that $\mathcal{U}^* = 0$ so that each of the three test statistics -- W, R, L -- is distributed as a non-central chi-square with non-centrality parameter α_n . Show that $\ell im_{n\to\infty}\alpha_n = \Delta' H'(H V H')^{-1} H\Delta/2$ in each case with $H = (\partial/\partial\lambda')h(\lambda^*)$ and $V = (\mathcal{J}^*)^{-1}\, \mathcal{I}^*(\mathcal{J}^*)^{-1}$.

8. Fix a realization of the errors. For large enough n, $\hat{\lambda}_n$ and $\tilde{\lambda}_n$ must be in an open neighborhood of λ^* on which $(\partial^2/\partial\lambda\partial\lambda')s_n(\lambda)$ is invertable. Why? Use Taylor's theorem to show that for large enough n, L is exactly given as a quadratic form in $(\partial/\partial\lambda)s_n(\tilde{\lambda}_n)$.

9. Using the identity derived in Section 6 verify the alternative form for α given in the examples following Theorems 14 and 15.

10. Verify the claim in Assumption 13 that $h(\lambda_n^o) = 0$ for all n implies that there is an N with $\lambda_n^o = \lambda_n^*$ for all $n > N$.

6. ALTERNATIVE REPRESENTATION OF A HYPOTHESIS

The results of the previous section presume that the hypothesis is stated as a parametric restriction

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

As we have seen, at times it is much more natural to express a hypothesis as a functional dependency

$$H: \lambda_n^\circ = g(\rho) \text{ for some } \rho \text{ in } \mathcal{R} \text{ against } A: \lambda_n^\circ \neq g(\rho) \text{ for any } \rho \text{ in } \mathcal{R} \ .$$

Suppose these two hypotheses are equivalent in the sense that there is a once differentiable function $\varphi(\lambda)$ defined on Λ such that the transformation

$$\tau = h(\lambda)$$
$$\rho = \varphi(\lambda)$$

has a once differentiable inverse

$$\lambda = \Psi(\rho, \tau)$$

with

$$g(\rho) = \Psi(\rho, 0) \ .$$

The set Λ is the set over which $s_n(\lambda)$ is to be minimized when computing $\hat{\lambda}_n$. Thus,

$$\mathcal{R} = \{\rho : \rho = \varphi(\lambda),\ h(\lambda) = 0,\ \lambda \text{ in } \Lambda\}$$

is the set over which $g(\rho)$ is defined; τ is a q-vector and ρ is an r-vector with $p = r + q$. To see that the existence of $\varphi(\lambda)$ implies that the two formulations of the null hypothesis are equivalent note that

$$\{\lambda : h(\lambda) = 0,\ \lambda \text{ in } \Lambda\}$$
$$= \{\lambda : \lambda = g(\rho),\ \rho \text{ in } \mathcal{R}\}\ .$$

Similarly, both formulations of the alternative hypotheses define the same set of admissible values for λ .

Since Theorem 11 is of little use without actually having $h(\lambda)$ at hand, we shall pass on to Theorems 14 and 15 and show that all the required computations can be performed knowing only $g(\rho)$. $\tilde{\lambda}_n$ can be computed by minimizing the composite function $s_n[g(\rho)]$ over $\mathcal{R}$ to obtain $\tilde{\rho}_n$ and putting $\tilde{\lambda}_n = g(\tilde{\rho}_n)$. Similarly, λ_n^* can be computed by minimizing $s_n^o[g(\rho)]$ over $\mathcal{R}$ to obtain ρ_n^o and putting $\lambda_n^* = g(\rho_n^o)$. The statistics R and L, the vector $(\partial/\partial\lambda)s_n^o(\lambda_n^*)$, and matrices $\mathcal{J}$, $\mathcal{J}$, $\mathcal{U}$ and V can now be computed directly. What remains is to compute matrices of the form $H'(HAH')^{-1}H$ where A is a computable, positive definite, symmetric matrix and $H = (\partial/\partial\lambda')h(\lambda_n^*)$. Let

$$G = (\partial/\partial\rho')g(\rho_n^o)\ .$$

We shall show that

$$H'(HAH')^{-1}H = A^{-1} - A^{-1}G(G'A^{-1}G)^{-1}G'A^{-1}$$

for any positive definite symmetric A.

By differentiating the equations

$$0 = h[g(\rho)]$$

$$\rho = \varphi[g(\rho)]$$

and evaluating the derivatives at $\rho = \rho_n^*$ we have

$$0 = HG$$

$$I = (\partial/\partial\lambda')\varphi(\lambda_n^*)G$$

which implies that rank (G) = r ; recall that rank (H) = q by assumption.

Factor A as $A = PP'$ (Problem 1 of Section 5). Trivially $HPP^{-1}G = 0$ which implies that there is a non-singular matrix B of order q and there is a non-singular matrix C of order r such that $\Theta_1 = P'H'B$ has orthonormal columns, $\Theta_2 = P^{-1}GC$ has orthonormal columns, and the matrix $\Theta = [\Theta_1 \vdots \Theta_2]$ is orthogonal. Then

$$\begin{aligned} I &= [\Theta_1 \vdots \Theta_2] \begin{bmatrix} \Theta_1' \\ \Theta_2' \end{bmatrix} \\ &= \Theta_1\Theta_1' + \Theta_2\Theta_2' \\ &= \Theta_1(\Theta_1'\Theta_1)^{-1}\Theta_1' + \Theta_2(\Theta_2'\Theta_2)^{-1}\Theta_2' \\ &= P'H'B(B'HPP'H'B)^{-1}B'HP \\ &\quad + P^{-1}GC[C'G'(P^{-1})'P^{-1}GC]^{-1}C'G'(P^{-1})' \\ &= P'H'(HAH')^{-1}HP + P^{-1}G(G'A^{-1}G)^{-1}G'(P^{-1})' \; . \end{aligned}$$

Whence

$$\begin{aligned} A^{-1} &= (P^{-1})'I(P^{-1}) \\ &= H'(HAH')^{-1}H + A^{-1}G(G'A^{-1}G)^{-1}G'A^{-1} \; . \end{aligned}$$

To illustrate, suppose that $\mathcal{I} = \mathcal{J}$ in Theorem 15. Then the non-centrality parameter is

$$\begin{aligned} \alpha &= n(\partial/\partial\lambda')s_n^{\circ}(\lambda_n^{*})\mathcal{J}^{-1}H'(H\mathcal{J}^{-1}H')^{-1}H\mathcal{J}^{-1}(\partial/\partial\lambda)s_n^{\circ}(\lambda_n^{*}) \\ &= n(\partial/\partial\lambda')s_n^{\circ}(\lambda_n^{*})[\mathcal{J}^{-1} - G(G'\mathcal{J}G)^{-1}G'](\partial/\partial\lambda)s_n^{\circ}(\lambda_n^{*}) \\ &= n(\partial/\partial\lambda')s_n^{\circ}(\lambda_n^{*})\mathcal{J}^{-1}(\partial/\partial\lambda)s_n^{\circ}(\lambda_n^{*}) \end{aligned}$$

since $(\partial/\partial\lambda)s_n^{\circ}(\lambda_n^{*}) = H'\theta$ where θ is the Lagrange multiplier.

7. RANDOM REGRESSORS

As noted earlier, the standard assumption in regression analysis is that the observed independent variables $\{x_t\}_{t=1}^n$ are fixed. With a model such as

$$y_t = g(x_t,\gamma_n^\circ) + e_t \qquad t = 1, 2, \ldots, n$$

the independent variables $\{x_t\}_{t=1}^n$ are held fixed and the sampling variation enters via sampling variation in the errors $\{e_t\}_{t=1}^n$. If the independent variables are random variables then the analysis is conditional on that realization $\{x_t\}_{t=1}^n$ that obtains. Stated differently, the model

$$y_t = g(x_t,\gamma_n^\circ) + e_t \qquad t = 1, 2, \ldots, n$$

defines the conditional distribution of $\{y_t\}_{t=1}^n$ given $\{x_t\}_{t=1}^n$ and the analysis is based on the conditional distribution.

An alternative approach is to assume that the independent variables $\{x_t\}_{t=1}^n$ are random and to allow sampling variation to enter both through the errors $\{e_t\}_{t=1}^n$ and the independent variables $\{x_t\}_{t=1}^n$. We shall see that the theory developed thus far is general enough to accomodate an assumption of random regressors and that the results are little changed save in one instance, that instance being the misspecified model. Therefore we shall focus the discussion on this case.

We have seen that under the fixed regressor setup the principal consequence of misspecification is the inability to estimate the matrix $\mathcal{J}^*$ from sample information because the obvious estimator $\hat{\mathcal{J}}$ converges almost surely to $\mathcal{J}^* + \mathcal{U}^*$ rather than to $\mathcal{J}^*$. As a result, test statistics are distributed asymptotically as general quadratic forms in normal random variables rather than as non-central chi-square random variables. In contrast, a consequence of the assumption of random regressors is that $\mathcal{U}^* = 0$. With random regressors test statistics are distributed asymptotically as the non-central chi-square. Considering least mean difference estimators, let us trace through the details as to why this is so. Throughout, $\mathcal{J} = \mathcal{J}_n^o$, $\mathcal{I} = \mathcal{I}_n^o$, and $\mathcal{U} = \mathcal{U}_n^o$.

With least mean distance estimators, the problem of non-zero $\mathcal{U}^*$ originates with the variables

$$(\partial/\partial\lambda)s[Y(e_t,x_t,\gamma_n^o),\ x_t,\tau_n^o,\lambda] \qquad t = 1,\ 2,\ \ldots,\ n$$

that appear in the proof of Theorem 4 . In a correctly specified situation, sensible estimation procedures will have the property that at each x the minimum of

$$\int_{\mathcal{E}} s[Y(e,x,\gamma_n^o),x,\tau_n^o,\lambda]dP(e)$$

will occur at $\lambda = \lambda_n^o$. Under the regularity conditions, this implies that

$$\begin{aligned} 0 &= (\partial/\partial\lambda)\int_{\mathcal{E}} s[Y(e,x,\gamma_n^o),x,\tau_n^o,\lambda_n^o]dP(e) \\ &= \int_{\mathcal{E}} (\partial/\partial\lambda)s[Y(e,x,\gamma_n^o),x,\tau_n^o,\lambda_n^o]dP(e) \ . \end{aligned}$$

Thus, the random variables

$$Z_t(e_t) = (\partial/\partial\lambda)s[Y(e_t,x_t,\gamma_n^o),x_t,\tau_n^o,\lambda_n^o]$$

have mean zero and their normalized sum

$$(1/\sqrt{n}\,)\Sigma_{t=1}^n Z_t(e_t)$$

has variance-covariance matrix

$$\mathcal{J} = (1/n)\Sigma_{t=1}^{n} \int_{\mathcal{E}} Z_t(e)Z_t'(e)dP(e)$$

which can be estimated by $\hat{\mathcal{J}}$. But in an incorrectly specified situation the mean of $Z_t(e_t)$ is

$$\mu_t = \int_{\mathcal{E}} (\partial/\partial\lambda)s[Y(e,x_t,\gamma_n^\circ),x_t,\tau_n^\circ,\lambda_n^\circ]dP(e) \; .$$

In general $\mu_t \neq 0$ and μ_t varies systematically with x_t . Under misspecification, the normalized sum

$$(1/\sqrt{n}\,)\,\Sigma_{t=1}^{n} Z_t$$

has variance-covariance matrix

$$\mathcal{J} = (1/n)\Sigma_{t=1}^{n}\int_{\mathcal{E}}[Z_t(e) - \mu_t][Z_t(e) - \mu_t]'dP(e)$$

as before but $\hat{\mathcal{J}}$ is, in essence, estimating

$$\mathcal{J} + \;\; (1/n)\Sigma_{t=1}^{n}\mu_t\mu_t' \; .$$

Short of assuming replicates at each point x_t , there seems to be no way to form an estimate of

$$\mathcal{U} = (1/n)\Sigma_{t=1}^{n}\mu_t\mu_t' \; .$$

Without being able to estimate $\mathcal{U}$, one cannot estimate $\mathcal{J}$.

The effect of an assumption of random regressors is to convert the deterministic variation in μ_t to random variation. The μ_t become independently distributed each having mean zero. From the point of view of the fixed regressors theory, one could argue that the independent variables have all been set to a constant value so that each observation is now a replicate. We illustrate with Example 3 and then return to the general discussion.

EXAMPLE 3 (Continued) To put the model into the form of a random regressors model within the framework of the general theory, let the data be generated according to the model

$$y_{(1)t} = g(y_{(2)t}, \gamma_n^o) + e_{(1)t} ,$$

$$y_{(2)t} = \mu_{(2)} + e_{(2)t} ,$$

which we presume satisfies Assumptions 1 through 3 with $x_t \equiv 1$ and μ the measure putting all its mass at $x = 1$; in other words, x_t enters the model trivially. The $y_{(2)t}$ are the random regressors. Convention has it, in this type of analysis, that $y_{(2)t}$ and $e_{(1)t}$ are independent whence $P(e)$ is a product measure

$$dP(e) = dP_{(1)}(e_{(1)}) \times dP_{(2)}(e_{(2)}) .$$

The fitted model is

$$y_{(1)t} = f(y_{(2)t},\lambda) + \mu_t \qquad t = 1, 2, \ldots, n$$

and λ is estimated by $\hat{\lambda}_n$ that minimizes

$$s_n(\lambda) = (1/n)\Sigma_{t=1}^n[y_{(1)t} - f(y_{(2)t},\lambda)]^2 .$$

Let ν be the measure defined by

$$\int_{\mathcal{Y}_{(2)}} g(y_{(2)})d\nu(y_{(2)}) = \int_{\mathcal{E}_{(2)}} g(\mu_{(2)} + e_{(2)})dP_{(2)}(e_{(2)})$$

where $\mathcal{Y}_{(2)}$ is the set of admissible values for the random variable $y_{(2)}$. We have:

$$s(y,x,\lambda) = [y_{(1)} - f(y_{(2)},\lambda)]^2$$

$$s_n^\circ(\lambda) = \sigma^2_{(1)} + \int_{\mathcal{Y}_{(2)}} [g(y_{(2)},\gamma_n^\circ) - f(y_{(2)},\lambda)]^2 d\nu(y_{(2)})$$

$$(\partial/\partial\lambda)s_n^\circ(\lambda) = -2\int_{\mathcal{Y}_{(2)}} [g(y_{(2)},\gamma_n^\circ) - f(y_{(2)},\lambda)]^2(\partial/\partial\lambda)f(y_{(2)},\lambda)d\nu(y_{(2)})$$

$$\lambda_n^\circ \text{ minimizes } \int_{\mathcal{Y}_{(2)}} [g(y_{(2)},\gamma_n^\circ) - f(y_{(2)},\lambda)]^2 d\nu(y_{(2)}) \quad .$$

The critical change from the fixed regressor case occurs in the computation of

$$\int_{\mathcal{E}} Z_t(e)dP(e) = \int_{\mathcal{E}} (\partial/\partial\lambda)s[Y(e,x_t,\gamma_n^\circ),x_t,\lambda_n^\circ]dP(e)\Big|_{x_t\equiv 1} \quad .$$

Let us decompose the computation into two steps. First compute the conditional mean of $Z_t(e)$ given that $y_{(2)} = y_{(2)t}$:

$$\mu_t = \int_{\mathcal{E}_{(1)}} Z_t(e_{(1)},e_{(2)})dP_{(1)}(e_{(1)})$$

$$= [g(y_{(2)t},\gamma_n^\circ) - f(y_{(2)t},\lambda_n^\circ)](\partial/\partial\lambda)f(y_{(2)t},\lambda_n^\circ) \quad .$$

Second compute the mean of μ_t

$$\int_{\mathcal{Y}_{(2)}} \mu_t d\nu(y_{(2)})$$

$$= \int_{\mathcal{Y}_{(2)}} [g(y_{(2)},\gamma_n^\circ) - f(y_{(2)},\lambda_n^\circ)](\partial/\partial\lambda)f(y_{(2)},\lambda_n^\circ)d\nu(y_{(2)})$$

$$= (\partial/\partial\lambda)s_n^\circ(\lambda_n^\circ)$$

$$= 0$$

because λ_n^o minimizes $s_n^o(\lambda)$. Consequently

$$\int_{\mathcal{E}} Z_t(e)dP(e) = 0$$

and $\mathscr{U} = \mathscr{U}^* = 0$. One can see that in the fixed regressor case the conditional mean μ_t of $s[Y(e,x,\gamma_n^o),x,\lambda_n^o]$ given the regressor is treated as deterministic quantity whereas in the random regressor case the conditional mean μ_t is treated as a random variable having mean zero.

Further computations yield:

$$\mathscr{I} = 4\sigma_{(1)}^2 \int_{\mathcal{Y}_{(2)}} (\partial/\partial\lambda)f(y_{(2)},\lambda_n^o)(\partial/\partial\lambda')f(y_{(2)},\lambda_n^o)d\nu(y_{(2)})$$

$$\mathscr{J} = 2\int_{\mathcal{Y}_{(2)}} (\partial/\partial\lambda)f(y_{(2)},\lambda_n^o)(\partial/\partial\lambda')f(y_{(2)},\lambda_n^o)d\nu(y_{(2)})$$

$$-2\int_{\mathcal{Y}_{(2)}} [g(y_{(2)},\gamma_n^o) - f(y_{(2)},\lambda_n^o)](\partial^2/\partial\lambda\partial\lambda')f(y_{(2)},\lambda_n^o)d\nu(y_{(2)}) \text{ . } \square$$

Returning to the general case, use the same strategy employed in the example to write

$$q(y,x,\gamma_n^\circ) = \begin{pmatrix} e_{(1)} \\ e_{(2)} \end{pmatrix} = \begin{pmatrix} q_{(1)}(y_{(1)},y_{(2)},\gamma_n^\circ) \\ y_{(2)} - \mu_{(2)} \end{pmatrix}$$

with $x \equiv 1$ and

$$dP(e) = dP_{(1)}(e_{(1)}) \times dP_{(2)}(e_{(2)}) \; ;$$

$y_{(2)}$ is the random regressor. The reduced form can be written as

$$y_{(1)} = Y_{(1)}(e_{(1)},y_{(2)},\gamma_n^\circ)$$

$$y_{(2)} = \mu_{(2)} + e_{(2)} \; .$$

Let ν be the measure such that

$$\int_{\mathcal{Y}_{(2)}} g(y_{(2)}) d\nu(y_{(2)}) = \int_{\mathcal{E}_{(2)}} g(\mu_{(2)} + e_{(2)}) dP_{(2)}(e_{(2)})$$

where $\mathcal{Y}_{(2)}$ is the set of admissible values of the random regressor $y_{(2)}$.

The distance function for a least mean distance estimator will have the form

$$s(y_{(1)},y_{(2)},\tau,\lambda) \; .$$

Since the distance function depends trivially on x_t, we have

$$s_n^\circ(\lambda_n^\circ) = \int_{\mathcal{E}} s[Y_{(1)}(e_{(1)},\mu_{(2)} + e_{(2)},\gamma_n^\circ), \mu_{(2)} + e_{(2)}, \tau_n^\circ,\lambda_n^\circ] dP(e)$$

$$= \int_{\mathcal{Y}_{(2)}} \int_{\mathcal{E}_{(1)}} s[Y_{(1)}(e_{(1)},y_{(2)},\gamma_n^\circ),y_{(2)},\tau_n^\circ,\lambda_n^\circ] \; dP_{(1)}(e_{(1)}) d\nu(y_{(2)}) \; .$$

Since $(\partial/\partial\lambda)s_n^\circ(\lambda_n^\circ) = 0$ and the regularity conditions permit interchange of differentiation and integration we have

$$\int_{\mathcal{Y}_{(2)}}\int_{\mathcal{E}_{(1)}} (\partial/\partial\lambda)s[Y_{(1)}(e_{(1)},y_{(2)},\gamma_n^\circ),y_{(2)},\tau_n^\circ,\lambda_n^\circ]dP_{(1)}(e_{(1)})d\nu(y_{(2)}) = 0$$

whence $\mathcal{U} = \mathcal{U}^* = 0$. Other computations assume a similar form, for example

$$\mathcal{J} = \int_{\mathcal{Y}_{(2)}}\int_{\mathcal{E}_{(1)}} (\partial^2/\partial\lambda\partial\lambda')s[Y_{(1)}(e_{(1)},y_{(2)},\gamma_n^\circ),y_{(2)},\tau_n^\circ,\lambda_n^\circ]dP_{(1)}(e_{(1)})d\nu(y_{(2)}) .$$

Sample quantities retain their previous form, for example

$$\mathcal{J} = (1/n)\Sigma_{t=1}^{n}(\partial^2/\partial\lambda\partial\lambda')s(y_{(1)t},y_{(2)t},\hat{\tau}_n,\hat{\lambda}_n) .$$

For a method of moments estimator, in typical cases one can exploit the structure of the problem and show directly that

$$\int_{\mathcal{Y}_{(2)}}\int_{\mathcal{E}_{(2)}} m[Y_{(1)}(e_{(1)},y_{(2)},\gamma_n^\circ),y_{(2)},\tau_n^\circ,\lambda_n^\circ]dP_{(1)}(e_{(1)})d\nu(y_{(2)}) = 0 .$$

This implies that $K = K^* = 0$ whence $\mathcal{U} = \mathcal{U}^* = 0$. The remaining computations are modified similarly to the foregoing.

8. CONSTRAINED ESTIMATION

Throughout we shall assume that the constraint has two equivalent representations:

$$\text{Parametric restriction: } h(\lambda) = 0\,, \quad \lambda \text{ in } \Lambda^*\,,$$

$$\text{Functional dependency: } \lambda = g(\rho)\,, \quad \rho \text{ in } \mathcal{R}\,,$$

where h: $R^p \to R^q$, g: $R^r \to R^p$; and $r + q = p$. They are equivalent in the sense that the null space of $h(\lambda)$ is the range space of $g(\rho)$:

$$\Lambda_H = \{\lambda\colon h(\lambda) = 0,\ \lambda \text{ in } \Lambda^*\} = \{\lambda\colon \lambda = g(\rho),\ \rho \text{ in } \mathcal{R}\}\ .$$

We also assume that both $g(\rho)$ and $h(\lambda)$ are twice continuously differentiable. From

$$h[g(\rho)] = 0$$

we have

$$(\partial/\partial\lambda')h[g(\rho)](\partial/\partial\rho')g(\rho) = HG = 0\ .$$

If rank $[H' \vdots G] = p$, we have from Section 6 that for any symmetric, positive definite matrix $\mathcal{L}$

$$G(G'\mathcal{L}G)^{-1}G' = \mathcal{L}^{-1} - \mathcal{L}^{-1}H'(H\mathcal{L}^{-1}H')^{-1}H\mathcal{L}^{-1}\ .$$

Section 6 gives a construction which lends plausibility to these assumptions.

Let the data generating model satisfy Assumptions 1 through 3. Let the objective function $s_n[g(\rho)]$ satisfy either Assumptions 4 through 6 or Assumptions 8 through 11. Let

$$\hat{\rho}_n \text{ minimize } s_n[g(\rho)] \ ,$$

$$\rho_n^{\circ} \text{ minimize } s_n^{\circ}[g(\rho)] \ ,$$

$$\rho^* \text{ minimize } s^*[g(\rho)] \ .$$

Then from either Theorem 3 or Theorem 7 we have that

$$\ell im_{n\to\infty}\, \hat{\rho}_n = \rho^* \text{ almost surely } ,$$

$$\ell im_{n\to\infty}\, \rho_n^{\circ} = \rho^* \text{ almost surely } ,$$

and from either Theorem 5 or Theorem 9 that

$$\sqrt{n}\, (\hat{\rho}_n - \rho_n^{\circ}) \xrightarrow{\mathcal{L}} N[\,0,\ (\mathscr{J}_\rho^*)^{-1}\ \mathscr{I}_\rho^*(\mathscr{J}_\rho^*)^{-1}] \ .$$

The matrices $\mathscr{I}_\rho^*$ and $\mathscr{J}_\rho^*$ are of order r by r and can be computed by direct application of Notation 2 or Notation 6 . In these computations one is working in an r-dimensional space not in a p-dimensional space. We emphasize this point with the ρ-subscript: $\mathscr{I}_\rho^*$, $\mathscr{J}_\rho^*$, and $\mathcal{U}_\rho^*$. To illustrate, computing according to Notation 2 one has:

$$\begin{aligned}
\mathcal{U}_\rho^* = \int_{\mathcal{X}}\int_{\mathcal{E}} &\{(\partial/\partial\rho)s[Y(e,x,\gamma^*),x,\tau^*,g(\rho^*)]\ dP(e)\} \\
&\times \{\int_{\mathcal{E}} (\partial/\partial\rho)s[Y(e,x,\gamma^*),x,\tau^*,g(\rho^*)]dP(e)\}'d\mu(x) \\
\mathscr{I}_\rho^* = \int_{\mathcal{X}}\int_{\mathcal{E}} &\{(\partial/\partial\rho)s[Y(e,x,\gamma^*),x,\tau^*,g(\rho^*)]\} \\
&\times \{(\partial/\partial\rho)s[Y(e,x,\gamma^*),x,\tau^*,g(\rho^*)]\}'dP(e)\ d\mu(x) - \mathcal{U}_\rho^* \\
\mathscr{J}_\rho^* = (\partial^2/\partial_\rho\partial_\rho')&s^*[g(\rho^*)] \ .
\end{aligned}$$

Estimators of $\mathcal{J}^*_\rho$ and $\mathcal{I}^*_\rho$ are computed according to Notation 4 or Notation 9. To illustrate, computing according to Notation 4 one has:

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

$$\hat{\mathcal{I}}_\rho = (1/n)\Sigma_{t=1}^{n}(\partial^2/\partial\rho\partial\rho')s[y_t,x_t,\hat{\tau}_n,g(\hat{\rho}_n)] \quad .$$

As to testing hypotheses, the theory of Section 5 applies directly. The computations according to Notation 3 or Notation 8 are similar to those illustrated above.

Often results reported in terms of

$$\tilde{\lambda}_n = g(\hat{\rho}_n)$$

are more meaningful than results reported in terms of $\hat{\rho}_n$. As an instance, one wants to show the effect of a restriction by presenting $\hat{\lambda}_n$ and its (estimated) standard errors together with $\tilde{\lambda}_n$ and its (estimated) standard errors in a tabular display. To do this, let

$$\lambda^*_n = g(\rho^\circ_n)$$

$$\lambda^\# = g(\rho^*_n) \quad .$$

By continuity of $g(\rho)$

$$\ell im_{n\to\infty}\ \tilde{\lambda}_n = \lambda^\# \text{ almost surely}$$

$$\ell im_{n\to\infty}\ \lambda^*_n = \lambda^\# \quad .$$

Note that $\lambda^\#$ is not equal to λ^* of either Section 3 or Section 4 unless the Pitman drift assumption is imposed. From the Taylor series expansion

$$g(\hat{\rho}_n) - g(\rho_n^\circ) = [G^* + o_s(1)]\sqrt{n}\,(\hat{\rho}_n - \rho_n^\circ)$$

where $G^* = (\partial/\partial\rho')g(\rho^*)$ we have that

$$\sqrt{n}\,(\tilde{\lambda}_n - \lambda_n^*) \xrightarrow{\mathcal{L}} N[0, G^*(\mathcal{J}_\rho^*)^{-1}\mathfrak{I}_\rho^*(\mathcal{J}_\rho^*)^{-1}G^{*\prime}] \;.$$

The variance-covariance matrix is estimated by

$$\hat{G}(\hat{\mathcal{J}}_\rho)^{-1}\,\hat{\mathfrak{I}}_\rho(\hat{\mathcal{J}}_\rho)^{-1}\hat{G}'$$

where $\hat{G} = (\partial/\partial\rho')g(\hat{\rho}_n)$. Let $\tilde{\theta}_n$ be the Lagrange multiplier for the minimization of $s_n(\lambda)$ subject to $h(\lambda) = 0$ and let

$$\tilde{\mathcal{J}} = (\partial^2/\partial\lambda\partial\lambda')[s_n(\tilde{\lambda}_n) + \tilde{\theta}_n' h(\tilde{\lambda}_n)] \;.$$

One can show that (Problem 1)

$$\hat{\mathcal{J}}_\rho = \hat{G}'\tilde{\mathcal{J}}\hat{G}$$

and using the chain rule with either Notation 4 or Notation 9 one finds that

$$\hat{\mathfrak{I}}_\rho = \hat{G}'\tilde{\mathfrak{I}}\hat{G}$$

where $\tilde{\mathfrak{I}} = \mathfrak{I}_n(\tilde{\lambda}_n)$. Thus

$$\hat{G}(\hat{\mathcal{J}}_\rho)^{-1}\,\hat{\mathfrak{I}}_\rho(\hat{\mathcal{J}}_\rho)\hat{G}' = \hat{G}(\hat{G}'\tilde{\mathcal{J}}\hat{G})^{-1}\hat{G}'\tilde{\mathfrak{I}}\hat{G}\,(\hat{G}'\tilde{\mathcal{J}}\hat{G})^{-1}\hat{G}' \;.$$

Using the identity given earlier on, one has

$$\hat{G}(\hat{\mathcal{J}}_\rho)^{-1}\,\hat{\mathfrak{I}}_\rho(\hat{\mathcal{J}}_\rho)^{-1}\hat{G}'$$

$$= \tilde{\mathcal{J}}^{-1}[I - \tilde{H}'(\tilde{H}\tilde{\mathcal{J}}^{-1}\tilde{H}')^{-1}\tilde{H}\tilde{\mathcal{J}}^{-1}]\tilde{\mathfrak{I}}[I - \tilde{\mathcal{J}}^{-1}\tilde{H}'(\tilde{H}\tilde{\mathcal{J}}^{-1}\tilde{H}')\tilde{H}]\tilde{\mathcal{J}}^{-1}$$

Where $\tilde{H} = (\partial/\partial\lambda')h(\tilde{\lambda}_n)$. The right hand side of this expression can be computed from knowledge of $s_n(\lambda)$ and $h(\lambda)$ alone.

Similarly, if

$$\lambda^{\#} \text{ minimizes } s^{*}(\lambda) \text{ subject to } h(\lambda) = 0$$

with Lagrange multipliers $\theta^{\#}$

$$G^{*}(\mathcal{J}_{\rho}^{*})^{-1} \mathcal{I}_{\rho}^{*}(\mathcal{J}_{\rho}^{*})^{-1} G^{*\prime}$$

$$= \mathcal{L}^{-1}[I - H'(H\mathcal{L}^{-1}H')^{-1}H\mathcal{L}^{-1}]\mathcal{I}[I - \mathcal{L}^{-1}H'(H\mathcal{L}^{-1}H')H]\mathcal{L}^{-1}$$

where

$$H = (\partial/\partial\lambda')h(\lambda^{\#})$$

$$\mathcal{L} = (\partial^2/\partial\lambda\partial\lambda')\,[s^{*}(\lambda^{\#}) + \theta^{\#\prime}h(\lambda^{\#})]$$

$$\mathcal{I} = \bar{\bar{\mathcal{I}}}(\lambda^{\#}) \;.$$

Under a Pitman drift, $\theta^{\#} = 0$ and the expression that one might expect from the proof of Theorem 13 obtains.

PROBLEMS

1. Show that the equation $h[g(\rho)] = 0$ implies

$$\Sigma_{\ell=1}^{p}(\partial/\partial\lambda_\ell)\; h_u[g(\rho)]\; (\partial^2/\partial\rho_i\partial\rho_j)g_\ell(\rho)$$

$$= - \Sigma_{\ell=1}^{p}\; \Sigma_{k=1}^{p}\; (\partial^2/\partial\lambda_k\partial\lambda_\ell)h_u[g(\rho)](\partial/\partial\rho_j)g_k(\rho)(\partial/\partial\rho_i)g_\ell(\rho).$$

Suppose that $\tilde{\lambda} = g(\hat{\rho})$ minimizes $s(\lambda)$ subject to $h(\lambda) = 0$ and that $\tilde{\theta}$ is the corresponding vector of Lagrange multipliers. Show that

$$\Sigma_{\ell=1}^{p}(\partial/\partial\lambda_\ell)s[g(\hat{\rho})](\partial^2/\partial\rho_i\partial\rho_j)g_\ell(\hat{\rho})$$

$$= \Sigma_{\ell=1}^{p}\; \Sigma_{k=1}^{p}\; \Sigma_{u=1}^{q}\; \tilde{\theta}_u(\partial^2/\partial\lambda_k\partial\lambda_\ell)h_u(\tilde{\lambda})\; (\partial/\partial\rho_j)g_k(\hat{\rho})(\partial/\partial\rho_i)g_\ell(\hat{\rho})\; .$$

Compute $(\partial^2/\partial\rho_i\partial\rho_j)s[g(\hat{\rho})]$ and substitute the expression above to obtain

$$(\partial^2/\partial\rho\partial\rho')s[g(\hat{\rho})] = [(\partial/\partial\rho')g(\hat{\rho})]'(\partial^2/\partial\lambda\partial\lambda')[s(\tilde{\lambda}) + \tilde{\theta}'h(\tilde{\lambda})][(\partial/\partial\rho')g(\hat{\rho})]\; .$$

9. REFERENCES

Amemiya, Takeshi (1974), "The nonlinear two-stage least squares estimator," Journal of Econometrics 2, 105-110.

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

Balet-Lawrence, Sonia (1975), Estimation of the parameters in an implicit model by minimizing the sum of absolute value of order p . Ph.D. Dissertation, North Carolina State University, Raleigh, NC.

Barnett, William A. (1976), "Maximum likelihood and iterated Aitken estimation of nonlinear systems of equations," Journal of the American Statistical Association 71, 354-360.

Bartle, Robert G. (1964), The Elements of Real Analysis. New York: John Wiley and Sons.

Burguete, Jose Francisco (1980), Asymptotic theory of instrumental variables in nonlinear regression. Ph.D. Dissertation, North Carolina State University, Raleigh, NC.

Burguete, Jose Francisco, A. Ronald Gallant, Geraldo Souza (1982), "On unification of the asymptotic theory of nonlinear econometric models," Econometric Reviews 1, 151-190.

Christensen, Laurits R., Dale W. Jorgenson, and Lawrence J. Lau (1975), "Transcendental Logarithmic Utility Functions," The American Economic Review 65, 367-383.

Chung, Kai Lai (1974), A Course in Probability, 2nd ed. New York: Academic Press.

Durbin, J. (1970), "Testing for serial correlation in least-squares regression when some of the regressions are lagged dependent variables," Econometrica 38, 410-429.

Edmunds, D. E. and V. B. Moscatelli (1977), "Fourier approximation and embeddings of Sobolev spaces," Dissertationes Mathematicae, CXLV.

Gallant, A. Ronald (1973), "Inference for nonlinear models," Institute of Statistics Mimeograph Series No. 875, North Carolina State University, Raleigh, NC.

Gallant, A. Ronald (1975a), "The power of the likelihood ratio test of location in nonlinear regression models," Journal of the American Statistical Association 70, 199-203.

Gallant, A. Ronald (1975b), "Testing a subset of the parameters of a nonlinear regression model," Journal of the American Statistical Association 70, 927-932.

Gallant, A. Ronald (1975c), "Seemingly unrelated nonlinear regressions," Journal of Econometrics 3, 35-50.

Gallant, A. Ronald (1976), "Confidence regions for the parameters of a nonlinear regression model," Institute of Statistics Mimeograph Series No. 1077, North Carolina State University, Raleigh, NC.

Gallant, A. Ronald (1977a), "Three-stage least squares estimation for a system of simultaneous nonlinear implicit equations," Journal of Econometrics 5, 71-88.

Gallant, A. Ronald (1977b), "Testing a nonlinear specification, a nonregular case," Journal of the American Statistical Association 72, 523-530.

Gallant, A. Ronald (1981), "On the bias in flexible functional forms and an essentially unbiased form: the Fourier flexible form," Journal of Econometrics 15, 211-245.

Gallant, A. Ronald and Alberto Holly (1980), "Statistical inference in an implicit, nonlinear, simultaneous equation model in the context of maximum likelihood estimation," Econometrica 48, 697-720.

Gallant, A. Ronald and Dale W. Jorgenson (1979), "Statistical inference for a system of simultaneous, nonlinear implicit equations in the context of instrumental variable estimation," Journal of Econometrics 11, 275-302.

Golub, Gene H. and V. Pereyra (1973), "The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate," SIAM Journal of Numerical Analysis 10, 413-432.

Grossman, W. (1976), "Robust nonlinear regression," in Johannes Gordesch and Peter Naeve, eds. Compstat 1976, Proceedings in Computational Statistics, Wien, Austria: Physica-Verlag.

Hartley, H. O. and A. Booker (1965), "Nonlinear least squares estimation," Annals of Mathematical Statistics 36, 638-650.

Holly, Alberto (1978), "Tests of nonlinear statistical hypotheses in multiple equation nonlinear models," Cashiers du Laboratoire d'Econometrie, Ecole Polytechnique, Paris.

Huber, Peter J. (1964), "Robust estimation of a location parameter," Annals Mathematical Statistics 35, 73-101.

Huber, Peter J. (1982), "Comment on "The Unification of the Asymptotic Theory of Nonlinear Econometric Models'," Econometric Reviews 1, 191-192.

Jennrich, Robert I. (1969), "Asymptotic properties of nonlinear least squares estimation," The Annals of Mathematical Statistics 40, 633-643.

Jorgenson, Dale W. and J. Laffont (1974), "Efficient estimation of nonlinear simultaneous equations with additive disturbances," Annals of Economic and Social Measurement 3, 615-640.

Loeve, Michel (1963), Probability Theory, 3rd ed. Princeton, NJ: D. Van Nostrand Co.

Luenberger, David G. (1969), Optimization by Vector Space Methods. New York: John Wiley and Sons.

Malinvaud, E. (1970a), "The consistency of nonlinear regressions," The Annals of Mathematical Statistics 41, 956-969.

Malinvaud, E. (1970b), Statistical Methods of Econometrics. Amsterdam: North-Holland. Chapter 9.

Phillips, Peter C. B. (1982), "Comment on 'The Unification of the Asymptotic Theory of Nonlinear Econometric Models'," Econometric Reviews 1, 193-199.

Rao, C. Radhakrishna (1973), Linear Statistical Inference and Its Applications, 2nd ed., New York: John Wiley and Sons.

Royden, H. L. (1963), Real Analysis. New York: MacMillan Company.

Ruskin, David M. (1978), M-Estimates of Nonlinear Regression Parameters and Their Jackknife Constructed Confidence Intervals. Ph.D. Dissertation, UCLA.

Searle, S. R. (1971), Linear Models. New York: John Wiley and Sons.

Serfling, Robert J. (1980), Approximation Theorems of Mathematical Statistics. New York: John Wiley and Sons.

Souza, Geraldo (1979), Statistical Inference in Nonlinear Models: A Pseudo Likelihood Approach. Ph.D. Dissertation, North Carolina State University, Raleigh, NC.

Souza, Geraldo and A. Ronald Gallant (1979), "Statistical inference based on M-estimators for the multivariate nonlinear regression model in implicit form," Institute of Statistics Mimeograph Series No. 1229, North Carolina State University, Raleigh, NC.

Tucker, Howard G. (1967), A Graduate Course in Probability. New York: Academic Press.

Varian, Hal R. (1978), Microeconomic Analysis. New York: W. W. Norton.

White, Halbert (1980), "Nonlinear regression on cross section data," Econometrica 48, 721-746.

White, Halbert (1982), "Comment on 'The Unification of the Asymptotic Theory of Nonlinear Econometric Models'," Econometric Reviews 1, 201-205.

Wouk, Arthur (1979), A Course of Applied Functional Analysis. New York: John Wiley and Sons.

10. INDEX TO CHAPTER 3.