TESTING A SUBSET OF THE PARAMETERS OF A NONLINEAR REGRESSION MODEL

by

A. R. GALLANT

Institute of Statistics
Mimeograph Series No. 943
Raleigh - August 1974

TESTING A SUBSET OF THE PARAMETERS OF A NONLINEAR REGRESSION MODEL

by

A. R. Gallant*

*Assistant Professor of Economics and Statistics, Institute of Statistics, North Carolina State University, Raleigh, North Carolina 27607.

ABSTRACT

The problem of testing for the location of a subset of the parameters entering the response function of a nonlinear regression model is considered. Two test statistics--the Likelihood Ratio Test statistic and a test derived from the asymptotic normality of the least squares estimator--for this problem are discussed and compared. The large sample distributions of these statistics are derived and Monte-Carlo power estimates are compared with power computed using the large sample distributions.

1. INTRODUCTION

This study is a continuation of the work reported in [4] which contains a brief review of the literature. Here testing for the location of a subset of the parameters entering the response function of a nonlinear regression model is considered rather than a joint test of all the parameters. This problem is more useful in applications but more tedious to deal with analytically.

Two test statistics are considered for this problem--the Likelihood Ratio Test statistic and a test statistic derived from the asymptotic normality of the least squares estimator. The tests based on these statistics are equivalent in linear regression but differ in nonlinear regression. Two approximating random variables with known small sample distributions are derived; they are asymptotically equivalent to these test statistics. (The mathematical detail is deferred to the Appendix.) A Monte-Carlo study is employed to verify that estimates of the power of these tests computed using the distribution function of the approximating random variables are sufficiently accurate for use in applications. Lastly, considerations governing a choice between the two tests are discussed.

A more formal description of the problem studied in this article is the following. We consider testing the hypothesis of location

$$H: \ \tau = \tau_o \qquad \text{against} \qquad A: \ \tau \neq \tau_o$$

at the α level of significance when the data are responses y_t to inputs x_t generated according to the nonlinear regression model

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

where the parameter θ has been partitioned according to

$$\theta' = (\rho', \tau') .$$

The form of the response function $f(x_t, \theta)$ is known. The unknown parameter θ is known to be contained in the parameter space Ω which is a subset of the p-dimensional reals and the first r coordinates ρ of the parameter $\theta' = (\rho', \tau')$ are regarded as nuisance parameters. The inputs x_t are contained in $\mathcal{X}$ which is a subset of the k-dimensional reals. The errors e_t are assumed independent and normally distributed with mean zero and unknown variance σ^2 .

An example of a problem fitting this description is that of testing whether selected treatment contrasts are zero in a completely randomized experimental design where some covariate w such as age is known to affect the experimental material by shifting its expectation exponentially. Such a model would be written

$$z_{ij} = u + t_i + \alpha e^{\beta w_{ij}} + \epsilon_{ij} \qquad (i = 1, 2, \ldots, I; \quad j = 1, 2, \ldots, J)$$

following the usual conventions. Assume for simplicity of exposition that $I=2$ and the hypothesis of interest is H: $t_1 = t_2$ against A: $t_1 \neq t_2$. This hypothesis suggests a reparameterization to the model

$$y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_4 e^{\theta_3 x_{3t}} + e_t \qquad (t = 1, 2, \ldots, n)$$

where

$$y_t = z_{1t}, \quad x_t' = (1, 1, w_{1t}), \quad e_t = \epsilon_{1t} \qquad (t = 1, 2, \ldots, J) ,$$

$$y_t = z_{2t}, \quad x_t' = (0, 1, w_{2t}), \quad e_t = \epsilon_{2t} \qquad (t = J+1, \ldots, 2J) ,$$

and

$$\theta' = (t_1 - t_2, u + t_2, \beta, \alpha) .$$

With this parameterization the hypothesis of interest becomes H: $\theta_1 = 0$ against A: $\theta_1 \neq 0$. The parameter τ is θ_1 and the remaining parameters $\rho' = (\theta_2, \theta_3, \theta_4)$ are the nuisance parameters. This example is used for the Monte-Carlo simulations.

2. NOTATION, ASSUMPTIONS, AND A PRELIMINARY THEOREM

The following notation will be useful in the remainder of the paper.

Notation: Given the regression model

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

where θ^* denotes the true but unknown value of θ , the observations

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

and the hypothesis of location

$$H: \tau = \tau_o \qquad \text{against} \qquad A: \tau \neq \tau_o$$

where θ has been partitioned according to

$$\theta' = (\rho', \tau') ,$$

we define:

$$y = (y_1, y_2, \ldots, y_n) \qquad (n \times 1) ,$$

$$f(\theta) = (f(x_1, \theta), f(x_2, \theta), \ldots, f(x_n, \theta))' \qquad (n \times 1) ,$$

$e = (e_1, e_2, \ldots, e_n)' \quad (n \times 1)$,

$F(\theta)$ = the $n \times p$ matrix with typical element $\frac{\partial}{\partial \theta_j} f(x_t, \theta)$,

$F_1(\theta)$ = the $n \times r$ matrix formed by deleting the last $p-r$ columns of $F(\theta)$,

$P = F(\theta^*)[F'(\theta^*)F(\theta^*)]^{-1}F'(\theta^*) \quad (n \times n)$,

$P_1 = F_1(\theta^*)[F_1'(\theta^*)F_1(\theta^*)]^{-1}F_1'(\theta^*) \quad (n \times n)$,

$P^{\perp} = I-P \quad (n \times n)$,

$P_1^{\perp} = I-P_1 \quad (n \times n)$,

$\hat{\theta}$ = the $p \times 1$ vector minimizing $\Sigma_{t=1}^n \{y_t - f(x_t, \theta)\}^2$ over Ω ,

$s^2 = (y - f(\hat{\theta}))'(y - f(\hat{\theta}))/(n-p)$,

$\tilde{\rho}$ = the $r \times 1$ vector minimizing $\Sigma_{t=1}^n \{y_t - f(x_t, (\rho, \tau_o))\}^2$ for (ρ, τ_o) in Ω ,

$\hat{\sigma}^2 = (y - f(\hat{\theta}))'(y - f(\hat{\theta}))/n$,

$\tilde{\sigma}^2 = (y - f(\tilde{\rho}, \tau_o))'(y - f(\tilde{\rho}, \tau_o))/n$,

$\delta_o(\rho) = f(\theta^*) - f(\rho, \tau_o) \quad (n \times 1)$,

ρ_o = the $r \times 1$ vector minimizing $\delta_o'(\rho)\delta_o(\rho)$ for (ρ, τ_o) in Ω ,

C_{22} = the $q \times q$ matrix formed by deleting the first r rows and columns of $(F'(\theta^*)F(\theta^*))^{-1}$,

$\hat{C}_{22}$ = the $q \times q$ matrix formed by deleting the first r rows and columns of $(F'(\hat{\theta})F(\hat{\theta}))^{-1}$,

U_2 = the $q \times 1$ vector obtained by deleting the first r rows of $[F'(\theta^*)F(\theta^*)]^{-1}F'(\theta^*)e$.

For simplicity we will write F for $F(\theta^*)$, F_1 for $F_1(\theta^*)$, and δ_o for $\delta_o(\rho_o)$. Also, the following vectors and subvectors are related according to the equalities $\hat{\theta} = (\hat{\rho}, \hat{\tau})$, $\theta^* = (\rho^*, \tau^*)$, and $\theta_o = (\rho_o, \tau_o)$.

In the remainder of the paper it will be convenient to use the following conventions to refer to various distribution and density functions.

Densities and Distributions. Let $g(t; v, \lambda)$ denote the non-central chi-squared density function with v degrees-freedom and non-centrality λ [5, p. 74] and let $G(t; v, \lambda)$ denote the corresponding distribution function. Let $n(t; \mu, \sigma^2)$ denote the normal density function with mean μ and variance σ^2 and let $N(t; \mu, \sigma^2)$ denote the corresponding distribution function. Let $F'(t; v_1, v_2, \lambda)$ denote the non-central F-distribution function with v_1 numerator degrees-freedom, v_2 denominator degrees-freedom and non-centrality λ [5, pp. 77-78]. The central F-distribution with corresponding degrees-freedom is denoted by $F(t; v_1, v_2)$. Define $H(x; v_1, v_2, \lambda_1, \lambda_2)$ to be the distribution function given by

$$0, \qquad x \le 1,\ \lambda_2 = 0,$$

$$\int_o^\infty G(t/[x-1] + 2x\lambda_2/[x-1]^2;\ v_2,\ \lambda_2/[x-1]^2)g(t;\ v_1,\ \lambda_1)dt, \qquad x < 1,\ \lambda_2 > 0,$$

$$\int_0^\infty N(-t;\ 2\lambda_2,\ 8\lambda_2)g(t;\ v_1,\ \lambda_1)dt, \qquad x=1,\ \lambda_2 > 0,$$

$$1 - \int_0^\infty G(t/[x-1] + 2x\lambda_2/[x-1]^2;\ v_2,\ \lambda_2/[x-1]^2)g(t;\ v_1,\ \lambda_1)dt, \quad x > 1 .$$

It is shown in [4] that

$$H(x;\ v_1,\ v_2,\ \lambda_1,\ 0) = F'(v_2(x-1)/v_1;\ v_1,\ v_2,\ \lambda_1) .$$

The following regularity conditions are taken from [4] and are repeated here for the readers' convenience.

Definition. Let $\mathfrak{A}$ be the Borel subsets of $\mathfrak{X}$ and $\{x_t\}_{t=1}^\infty$ be the sequence of inputs chosen from $\mathfrak{X}$. Let $I_A(x)$ be the indicator function of a subset A of $\mathfrak{X}$. The measure μ_n on $(\mathfrak{X},\mathfrak{A})$ is defined by

$$\mu_n(A) = n^{-1} \Sigma_{t=1}^n I_A(x_t)$$

$$= \text{the proportion of } x_t \text{ with } t \le n \text{ which are in } A$$

for each $A \in \mathfrak{A}$.

Definition. A sequence of measures $\{\mu_n\}$ on $(\mathfrak{X},\mathfrak{A})$ is said to converge weakly to a measure μ on $(\mathfrak{X},\mathfrak{A})$ if for every real valued, bounded, continuous function g with domain $\mathfrak{X}$

$$\int g(x)d\mu_n(x) \rightarrow \int g(x)d\mu(x)$$

as $n \rightarrow \infty$.

Assumptions. The parameter space Ω and the set $\mathfrak{X}$ are compact subsets of the p-dimensional and k-dimensional reals, respectively. The response function $f(x,\theta)$ and the partial derivatives $\frac{\partial}{\partial\theta_i} f(x,\theta)$ and $\frac{\partial^2}{\partial\theta_i\partial\theta_j} f(x,\theta)$ are continuous on $\mathfrak{X} \times \Omega$. The sequence of inputs $\{x_t\}_{t=1}^{\infty}$ are chosen such that the sequence of measures $\{\mu_n\}_{n=1}^{\infty}$ converges weakly to a measure μ defined over $(\mathfrak{X},\mathfrak{a})$. The true value of θ, denoted by θ^*, is contained in an open set which, in turn, is contained in Ω. If $f(x,\theta) = f(x,\theta^*)$ except on a set of μ measure zero, it is assumed that $\theta = \theta^*$. The $p \times p$ matrix

$$\mathfrak{F} = [\int \frac{\partial}{\partial\theta_i} f(x,\theta^*) \frac{\partial}{\partial\theta_j} f(x,\theta^*) d\mu(x)]$$

is non-singular. As mentioned earlier, the errors $\{e_t\}$ are independent with density $n\{x;\ 0,\ \sigma^2\}$ where σ^2 is non-zero, finite, and unknown.

An additional assumption is used to obtain the characterizations of the test statistics studied in this paper. See [7, p. 305] for its motivation and justification.

Assumptions (continued). As n increases, τ_o tends to τ^* at a rate such that $\sqrt{n}\,(\tau_o - \tau^*)$ and $\sqrt{n}\,(\rho_o - \rho^*)$ converge to finite limits.

Theorem 1 gives the asymptotic properties of $\hat{\theta}$, $\hat{\sigma}^2$, and $\tilde{\sigma}^2$ which are used to characterize the two test statistics studied in this paper. This theorem is proved in the appendix.

Theorem 1. Let the assumptions listed above be satisfied.

The random variable $\hat{\theta}$ converges almost surely to θ^* and may be characterized as

$$\hat{\theta} = \theta^* + (F'F)^{-1}F'e + \alpha_n$$

where $\sqrt{n}\,\alpha_n$ converges in probability to zero.

The random variable $\hat{\sigma}^2$ converges almost surely to σ^2 and its reciprocal may be characterized as

$$1/\hat{\sigma}^2 = n/e'P^{\perp}e + a_n$$

where $n\,a_n$ converges in probability to zero.

The random variable $\tilde{\sigma}^2$ may be characterized as

$$\tilde{\sigma}^2 = (e + \delta_o)'P_1^{\perp}(e + \delta_o)/n + b_n$$

where $n\,b_n$ converges in probability to zero.

3. THE LIKELIHOOD RATIO TEST AND ITS LARGE SAMPLE DISTRIBUTION

The Likelihood of the sample y is

$$L(y;\, \theta,\, \sigma^2) = (2\pi\sigma^2)^{-n/2}\exp[-(1/2)(y - f(\theta))'(y - f(\theta))/\sigma^2] \; .$$

The maximum of the Likelihood subject to H is

$$L(H) = (2\pi\tilde{\sigma}^2)^{-n/2}\exp[-n/2]$$

and its maximum over the entire parameter space is

$$L(\Omega) = (2\pi\hat{\sigma}^2)^{-n/2}\exp[-n/2] \; .$$

Thus, the Likelihood Ratio is $L(H)/L(\Omega) = (\tilde{\sigma}^2/\hat{\sigma}^2)^{-n/2}$ and the Likelihood

Ratio Test has the form: Reject the null hypothesis H when the statistic

$$T = \tilde{\sigma}^2/\hat{\sigma}^2$$

is larger than c where $P[T > c \mid H] = \alpha$.

The computation of the statistic requires that two nonlinear models be fitted to the data using, say, either Hartley's [6] or Marquardt's [9] algorithm. The constrained model

$$y_t = f(x_t, (\rho, \tau_o)) + e_t$$

must be fit to obtain $\tilde{\sigma}^2$ and the unconstrained model

$$y_t = f(x_t, \theta) + e_t$$

to obtain $\hat{\sigma}^2$.

The statistical behavior of T is given by Theorem 2 which is proved in the Appendix.

<u>Theorem 2</u>. Under the assumptions listed in Section 2, the Likelihood Ratio Test statistic may be characterized as

$$T = X + c_n$$

where $n\, c_n$ converges in probability to zero and

$$X = (e + \delta_o)' P_1^{\perp} (e + \delta_o) / e' P^{\perp} e \ .$$

The random variable X has the distribution function $H(x;\, q,\, n\text{-}p,\, \lambda_1, \lambda_2)$ where $\lambda_1 = \delta_o'(P-P_1)\delta_o/(2\sigma^2)$ and $\lambda_2 = \delta_o' P^{\perp} \delta_o/(2\sigma^2)$.

A relationship between the distribution function of the approximating random variable X and the non-central F-distribution was mentioned in the previous section. Using this relationship, the critical point c^* of X $(P[X > c^* | \lambda_1 = \lambda_2 = 0] = \alpha)$ can be obtained from the formula

$$c^* = 1 + q\, F_\alpha/(n-p)$$

where F_α denotes the upper $\alpha \cdot 100$ percentage point of an F-distribution with q numerator degrees of freedom and $n-p$ denominator degrees of freedom. This critical point c^* may be used to approximate the critical point c of T -- in applications one rejects $H\colon\ \tau = \tau_0$ when the calculated value of T exceeds c^* .

The calculation of the non-centralities λ_1 and λ_2 for trial values of θ^* and σ^2 requires the minimization of the sum of squares $\Sigma_{t=1}^{n}[f(x_t, (\rho, \tau_0)) - f(x_t, \theta^*)]^2$ to obtain ρ_0 for computing δ_0 . This may be done using either Hartley's [6] or Marquardt's [9] algorithm with $f(\theta^*)$ replacing y . The remaining computations are straightforward matrix algebra.

A series expansion of the distribution function $H(x; v_1, v_2, \lambda_1, \lambda_2)$ is given in [4] which may be used to compute $P[X > c^*]$ once λ_1 and λ_2 have been obtained. However, in most applications λ_2 will be small $(< .1)$ and $P[X > c^*]$ can be adequately approximated by charts of the non-central F-distribution [1, 10]. The correspondence for the use of these charts is

$$H(c^*; v_1, v_2, \lambda_1, \lambda_2) \doteq F'(F_\alpha; v_1, v_2, \lambda_1) \ .$$

That is, the charts are entered at the α level of significance with q

numerator degrees of freedom, n-p denominator degrees of freedom, and non-centrality $\lambda_1 = \delta_o'(P-P_1)\delta_o/(2\sigma^2)$.

A Monte-Carlo study reported in [4] for an exponential model with $q = p = 2$ and $n = 30$ suggests that the approximation

$$P[T > c^*] \doteq P[X > c^*]$$

will be fairly accurate in applications. Additional evidence is given in Table 1. Using the example of Section 1, five thousand trials were generated with the parameters set at $n = 30$, $\theta_2 = 1$, $\theta_4 = -.5$ and $\sigma^2 = .001$ with θ_1 and θ_3 varied as shown in the table. The standard errors shown in the table refer to the fact that the Monte-Carlo estimator of $P[T > c^*]$ is binomially distributed with variance estimated by $P[X > c^*] \cdot P[X \leq c^*]/5000$.

- -

Table 1 about here

- -

4. A TEST STATISTIC DERIVED FROM THE ASYMPTOTIC NORMALITY OF THE LEAST SQUARES ESTIMATOR AND ITS LARGE SAMPLE DISTRIBUTION

The least squares estimator $\hat{\theta}$ is asymptotically normally distributed with mean θ^* and a variance-covariance matrix estimated consistently by $[F'(\hat{\theta})F(\hat{\theta})]^{-1} s^2$ under the assumptions of Section 2 [2, 3].1/ This suggests the use of the test statistic

$$S = \frac{(\hat{\tau} - \tau_o)'\hat{C}_{22}^{-1}(\hat{\tau} - \tau_o)/q}{s^2} .$$

The test has the form: Reject the null hypothesis H when S exceeds d where $P[S > d|H] = \alpha$.

The program used to compute the least squares estimator will usually print $\hat{\theta}$, $[F'(\hat{\theta})F(\hat{\theta})]^{-1}$, and s^2. If $q=1$, the test statistic can easily be computed from the printed results. However, if accuracy is to be maintained, some programming effort will be required to: recover $\hat{\theta}$, $\hat{C}_{22}$, and s^2, invert $\hat{C}_{22}$, and perform the necessary matrix algebra.

The statistical behavior of the statistic S is given by Theorem 3 which is proved in the Appendix.

<u>Theorem 3</u>. Under the assumptions listed in Section 2, the test statistic based on the asymptotic normality of the least squares estimator may be characterized as

$$S = Y + Y_n$$

where Y_n converges in probability to zero and

$$Y = \frac{(U_2 + \tau^* - \tau_o)' C_{22}^{-1} (U_2 + \tau^* - \tau_o)/(q)}{e' P^{\perp} e/(n-p)}$$

which is distributed as $F'(y;\, q,\, n-p,\, \lambda)$ where $\lambda = (\tau^* - \tau_o)' \times C_{22}^{-1}(\tau^* - \tau_o)/(2\sigma^2)$.

The critical point d^* of $Y(P[Y > d^* | \lambda=0] = \alpha)$ is obtained directly from a table of the F-distribution by setting $d^* = F_\alpha$ where F_α denotes the upper $\alpha \cdot 100$ percentage point with q numerator degrees of freedom and $n-p$ denominator degrees of freedom. This critical point may be used to approximate the critical point d of S -- in applications one rejects $H\colon\ \tau = \tau_o$ when the calculated value of S exceeds F_α .

Monte-Carlo evidence, Table 2, suggests that, under the alternative, the approximation

$$P[S > d^*] \doteq P[X > d^*]$$

is not as accurate as the approximation of T by X. However, the approximation appears to be sufficiently accurate for use in applications. Table 2 was constructed in the same manner as Table 1.

- -

Table 2 about here

- -

5. COMPARISON

Tests based on S and T are equivalent in linear regression; they are not in nonlinear regression. This raises the question of which should be used in applications.

Approaching this question on the basis of power, the analytical and Monte-Carlo evidence suggests that when a component of the parameter θ enters the model nonlinearly the Likelihood Ratio Test with respect to a hypothesis concerning it has better power than the test based on the asymptotic normality of the least squares estimator. When a component enters the model linearly the two tests have the same power. However, the evidence presented here is too limited to place much reliance on this generalization. If the question is of critical importance in an application, the non-centralities λ_1 and λ can be computed and compared. Provided λ_2 is small $(< .1)$ the Likelihood Ratio Test will have better power than the test based on the asymptotic normality of the least squares estimator when λ_1 is larger than λ. It is not necessary to evaluate $P[X > c^*]$ or $P[Y > d^*]$ if relative power is the only matter of concern; the comparison of non-centralities will suffice.

As regards computational convenience, we have seen previously that for q=1 the test based on the asymptotic normality is more convenient; for q > 1 the Likelihood Ratio Test is more convenient. One qualification is necessary. Hartley's [6] and Marquardt's [9] algorithms do not always converge in practice. Thus, it is possible that $\hat{\theta}$ can be successfully computed but not $\tilde{\rho}$.[2/] In such situations, $\tilde{\rho}$ must be computed by other means, e.g., grid search, or the test based on the asymptotic normality of the least squares estimator employed by default.

In most applications, a hypothesis involves only a single component of θ and, therefore, q=1. The author's practice is to use the test based on the asymptotic normality of the least squares estimator for computational convenience when the purpose is to assist a judgment as to whether the model employed adequately represents the data. He uses the Likelihood Ratio Test when the hypothesis is an important aspect of the study.

FOOTNOTES

1. $\sqrt{n}\,(\hat{\theta} - \theta^{*}) \xrightarrow{\mathcal{L}} N(0, \Sigma^{-1}\sigma^{2})$ and $[\frac{1}{n} F'(\hat{\theta})F(\hat{\theta})]^{-1} \xrightarrow{a.s.} \Sigma^{-1}\sigma^{2}$.

2. $\hat{\rho}$ is usually a good start value for computing $\tilde{\rho}$.

1. MONTE-CARLO POWER ESTIMATES FOR THE LIKELIHOOD RATIO TEST

Parameters		H: $\theta_1 = 0$ against A: $\theta_1 \neq 0$					H: $\theta_3 = -1$ against A: $\theta_3 \neq -1$				
					Monte-Carlo					Monte-Carlo	
θ_1	θ_3	λ_1	λ_2	$P[X > c^*]$	$P[T > c^*]$	STD. ERR.	λ_1	λ_2	$P[X > c^*]$	$P[T > c^*]$	STD. ERR.
0.0	-1.0	0.0	0.0	.050	.050	.003	0.0	0.0	.050	.052	.003
0.008	-1.1	0.2353	0.0000	.101	.094	.004	0.2421	0.0006	.103	.110	.004
0.015	-1.2	0.8307	0.0000	.237	.231	.006	0.8503	0.0078	.243	.248	.006
0.030	-1.4	3.3343	0.0000	.700	.687	.006	2.6669	0.0728	.618	.627	.007

2. MONTE-CARLO POWER ESTIMATES FOR THE TEST BASED ON THE ASYMPTOTIC NORMALITY OF THE LEAST SQUARES ESTIMATOR

Parameters		H: $\theta_1 = 0$ against A: $\theta_1 \neq 0$		Monte-Carlo		H: $\theta_3 = -1$ against A: $\theta_3 \neq -1$		Monte-Carlo	
θ_1	θ_3	λ	$P[Y > F_\alpha]$	$P[S > F_\alpha]$	STD. ERR.	λ	$P[Y > F_\alpha]$	$P[S > F_\alpha]$	STD. ERR.
0.0	-1.0	0.0	.050	.050	.003	0.0	050	.056	.003
0.008	-1.1	0.2353	.101	.094	.004	0.2220	.098	.082	.004
0.015	-1.2	0.8309	.237	.231	.006	0.7332	.215	.183	.006
0.030	-1.4	3.3343	.700	.687	.006	2.1302	.511	.513	.007

REFERENCES

[1] Fox, M., "Charts of the Power of the F-test," The Annals of Mathematical Statistics, 27 (June 1956), 484-97.

[2] Gallant, A. R., "Statistical Inference for Nonlinear Regression Models," Ph.D. dissertation, Iowa State University, 1971.

[3] Gallant, A. R., "Inference for Nonlinear Models," Institute of Statistics Mimeograph Series, No. 875. Raleigh: North Carolina State University, 1973.

[4] Gallant, A. R., "The Power of the Likelihood Ratio Test of Location in Nonlinear Regression Models," Journal of the American Statistical Association, 70 (in press).

[5] Graybill, F. A., An Introduction to Linear Statistical Models, Volume 1. New York: McGraw-Hill Book Company, 1961.

[6] Hartley, H. O., "The Modified Gauss-Newton Method for the Fitting of Non-linear Regression Functions by Least Squares," Technometrics, 3 (May 1961), 269-80.

[7] Lehmann, E. L., Testing Statistical Hypotheses. New York: John Wiley and Sons, Inc. 1959.

[8] Malinvaud, E., "The Consistency of Nonlinear Regressions," The Annals of Mathematical Statistics, 41 (June 1970), 956-69.

[9] Marquardt, D. W., "An Algorithm for Least-squares Estimation of Nonlinear Parameters," Journal of Industrial and Applied Mathematics, 11 (June 1963), 431-41.

[10] Pearson, E. S. and Hartley, H. O., "Charts of the Power Function of the Analysis of Variance Tests, Derived from the Non-central F-distribution," Biometrika, 38 (June 1951), 112-30.

APPENDIX. PROOFS OF THEOREMS

Three technical lemmas are needed in the proofs of the theorems.

Lemma 1. Let $g(x, \lambda)$ be a real valued continuous function defined on $\mathcal{X} \times \Lambda$ where Λ is a compact set. Under the assumptions listed in Section 2:

i) The series $\frac{1}{n} \Sigma_{t=1}^{n} g(x_t, \lambda)$ converges to the integral $\int g(x, \lambda)d\mu(x)$ uniformly in λ .

ii) The random variable $\frac{1}{n} \Sigma_{t=1}^{n} g(x_t, \lambda)e_t$ converges almost surely to zero uniformly in λ .

iii) If $\hat{\lambda}_n$ converges almost surely to λ^* in Λ then $\frac{1}{n} \Sigma_{t=1}^{n} g(x_t, \hat{\lambda}_n)$ converges almost surely to $\int g(x, \lambda^*)d\mu(x)$.

iv) If $\hat{\lambda}_n$ converges in probability to λ^* in Λ then $\frac{1}{n} \Sigma_{t=1}^{n} g(x_t, \hat{\lambda}_n)$ converges in probability to $\int g(x, \lambda^*)d\mu(x)$.

Proof. Part i is proved in [8, p. 967]. Part ii is proved as Lemma 3.4 in [3]. Parts iii and iv follow from the uniform convergence given in Part i. □

Lemma 2. Under the assumptions listed in Section 2, the random variable $\tilde{\rho}$ converges almost surely to ρ^* .

Proof. Let $Q_n(\theta) = (y - f(\theta))'(y - f(\theta))/n = e'e/n + 2e'(f(\theta^*) - f(\theta))/n + (f(\theta^*) - f(\theta))'(f(\theta^*) - f(\theta))/n$ which converges almost surely to $\bar{Q}(\theta) = \sigma^2 + \int \{f(x, \theta^*) - f(x, \theta)\}^2 d\mu(x)$ uniformly in θ by the Strong Law of Large Numbers and Lemma 1 i, ii. Let

$\{\tilde{\rho}_n\}$ be a sequence of points minimizing $Q_n(\rho, \tau_o)$ corresponding to a realization of the errors $\{e_t\}$. Since Ω is compact there is at least one limit point $\rho^{\#}$ and at least one subsequence $\{\tilde{\rho}_{n_m}\}$ such that $\lim_{m \to \infty} \tilde{\rho}_{n_m} = \rho^{\#}$. Unless the realization $\{e_t\}$ belongs to the exceptional set, $\sigma^2 \le \bar{Q}(\rho^{\#}, \tau^*) = \lim_{m \to \infty} Q(\tilde{\rho}_{n_m}, \tau_o) \le \lim_{m \to \infty} Q(\rho^*, \tau_o) = \bar{Q}(\theta^*) = \sigma^2$ so that $\int \{f(x, \theta^*) - f(x, (\rho^{\#}, \tau^*))\}^2 d\mu(x) = 0$ which implies $\rho^{\#} = \rho^*$ by assumption. Thus, excepting realizations in the exceptional set, $\tilde{\rho}_n$ has only one limit point which is ρ^*. □

Lemma 3. Under the assumptions listed in Section 2, the random vector $(1/\sqrt{n})F'e$ converges in distribution to a p-variate normal and $(1/\sqrt{n})F'\delta_o$ converges to a finite limit.

Proof. By Lemma 3.5 of [3] or Lemma 4.2 of [2] $(1/\sqrt{n})F'e$ converges in distribution to a p-variate normal. By Taylor's theorem, $(1/\sqrt{n})F'\delta_o = -(1/\sqrt{n})F'F(\bar{\theta})(\theta_o - \theta^*)$ where $\bar{\theta}$ is on the line segment joining θ_o to θ^*. The assumption that $\sqrt{n}\,(\theta_o - \theta^*)$ tends to a finite limit and Lemma 1, i applied to $\frac{1}{n} F'F(\bar{\theta})$ imply the result. □

Proof of Theorem 1. The first conclusion is proved as Theorem 3 of [3] or as Proposition 6.1 of [2]. The second conclusion is proved as Lemma 1 of [4].

By Lemma 2, $(\tilde{\rho}, \tau_o)$ will almost surely be contained in the open subset of Ω containing θ^* allowing the use of Taylor's expansions in the proof and causing $\tilde{\rho}$ to eventually become a stationary point of $Q(\rho, \tau_o) = (y - f(\rho, \tau_o))'(y - f(\rho, \tau_o))/n$.

In this paragraph we will obtain intermediate results based on Taylor's expansions which will be used later in the proof. Using Taylor's theorem,

$$f(\tilde{\rho}, \tau_o) = f(\theta_o) + F_1(\tilde{\rho} - \rho_o) + H(\tilde{\rho} - \rho_o)$$

where H is the $n \times r$ matrix with typical row $\nabla'_\rho f(x_t, \theta_o) - \nabla'_\rho f(x_t, \theta^*) + \frac{1}{2}(\tilde{\rho} - \rho_o)'\nabla^2_\rho f(x_t, (\bar{\rho}, \tau_o))$ and $\bar{\rho}$ is on the line segment joining $\tilde{\rho}$ to ρ_o. Using Lemmas 1 and 2 and the assumption that $\sqrt{n}\,(\theta_o - \theta^*)$ converges to a finite limit one can show that $\frac{1}{n} F_1'(\tilde{\rho}, \tau_o)H$, and $\frac{1}{n} F_1'H$, and $\frac{1}{n} H'H$ converge almost surely to the zero matrix. Applying Taylor's theorem twice one can write

$$[F_1'(\rho, \tau_o) - F_1'](e + \delta_o) = E(\theta_o - \theta^*) + D(\rho - \rho_o)$$

where E is the $r \times p$ matrix with typical element

$$e_{ij} = \Sigma_{t=1}^n \frac{\partial^2}{\partial\theta_j\partial\rho_i} f(x_t, \bar{\theta})[e_t + f(x_t, \theta^*) - f(x_t, \theta_o)]$$

and D is the $r \times r$ matrix with typical element

$$d_{ij} = \Sigma_{t=1}^n \frac{\partial^2}{\partial\rho_j\partial\rho_i} f(x_t, (\bar{\rho}, \tau_o))[e_t + f(x_t, \theta^*) - f(x_t, \theta_o)] .$$

We will write E_o when this matrix depends on ρ_o and will write $\tilde{E}$ and $\tilde{D}$ when the dependence is on $\tilde{\rho}$. Using Lemma 1 and the assumed convergence of θ_o to θ^* one can show that $\frac{1}{n} E_o$, $\frac{1}{n}\tilde{E}$, and $\frac{1}{n}\tilde{D}$ converge almost surely to the zero matrix.

In this paragraph we will obtain the probability order of $\tilde{\rho}$. As mentioned earlier, for almost every realization of the errors $\{e_t\}$, $\tilde{\rho}$ is eventually a stationary point of $(-\sqrt{n}/2)Q(\rho, \tau_o)$ so that the random vector $(-\sqrt{n}/2)\nabla_\rho Q(\tilde{\rho}, \tau_o) = (1/\sqrt{n})F_1'(\tilde{\rho}, \tau_o)(y - f(\tilde{\rho}, \tau_o))$ converges almost surely to the zero vector. Substituting the expansions in the previous paragraph we have that

$$[\frac{1}{n} F_1'(\tilde{\rho}, \tau_o)F_1 + \frac{1}{n} F_1'(\tilde{\rho}, \tau_o)H - \frac{1}{n}\tilde{D}]\sqrt{n}(\tilde{\rho}-\rho_o) - \frac{1}{\sqrt{n}}\tilde{E}(\theta_o-\theta^*) - \frac{1}{\sqrt{n}}F_1'(e+\delta_o)$$

converges almost surely to zero. Lemma 1 and our previous results imply that the matrix in brackets converges almost surely to a positive definite submatrix of $\mathfrak{T}$. Lemma 3 implies that $(1/\sqrt{n})F_1'(e + \delta_o)$ converges in distribution to a r-variate normal; $(\frac{1}{n}\tilde{E})\sqrt{n}(\theta_o - \theta^*)$ converges almost surely to the zero vector. These facts allow us to conclude that

$$u_n = \sqrt{n}\,[\tilde{\rho} - \rho_o - (F_1'F_1)^{-1}F_1'(e + \delta_o)]$$

converges in probability to zero and that $v_n = \sqrt{n}(\tilde{\rho} - \rho_o)$ is bounded in probability.

The sum of squares

$$\|y - f(\tilde{\rho},\tau_o)\|^2 = \|e + \delta_o + f(\theta_o) - f(\tilde{\rho},\tau_o)\|^2$$

$$= \|P_1^\perp(e+\delta_o) + P_1(e+\delta_o) - F_1(\tilde{\rho}-\rho_o) - H(\tilde{\rho}-\rho_o)\|^2$$

$$= \|P_1^\perp(e+\delta_o)\|^2 - 2(e+\delta_o)'P_1^\perp H(\tilde{\rho}-\rho_o)$$

$$+ \|P_1(e+\delta_o) - F_1(\tilde{\rho}-\rho_o) - H(\tilde{\rho}-\rho_o)\|^2 \quad .$$

The cross product term may be written as

$$(e+\delta_o)'P_1^{\perp}H(\tilde{\rho}-\rho_o) = (e+\delta_o)'H(\tilde{\rho}-\rho_o) - (e+\delta_o)'F_1(F_1'F_1)^{-1}F_1'H(\tilde{\rho}-\rho_o)$$

$$= \sqrt{n}\,(\theta_o-\theta^*)'(\tfrac{1}{n}\,E_o)'v_n$$

$$+ \tfrac{1}{2}\,v_n'(\tfrac{1}{n}\,\Sigma_{t=1}^{n}[e_t + f(x_t,\theta^*) - f(x_t,\theta_o)]\nabla_\rho^2 f(x_t,\,(\bar{\rho},\tau_o))v_n$$

$$- \frac{1}{\sqrt{n}}\,(e+\delta_o)'F_1(\tfrac{1}{n}\,F_1'F_1)^{-1}(\tfrac{1}{n}\,F_1'H)v_n \;.$$

Each of these terms converges almost surely to zero by Lemmas 1, 2, and 3 and our previous results. By the triangle inequality

$$\|P_1(e+\delta_o) + F_1(\tilde{\rho}-\rho_o) - H(\tilde{\rho}-\rho_o)\| \le \|F_1[(F_1'F_1)^{-1}F_1'(e+\delta) - (\tilde{\rho}-\rho_o)]\|$$

$$+ \|H(\tilde{\rho}-\rho_o)\|$$

$$= (u_n'(\tfrac{1}{n}\,F_1'F_1)u_n)^{\frac{1}{2}}$$

$$+ (v_n'(\tfrac{1}{n}\,H'H)v_n)^{\frac{1}{2}} \quad .$$

The two terms on the right converge in probability to zero by our previous results. □

<u>Proof of Theorem 2</u>. Applying Theorem 1 we have

$$T = X + a_n b_n + (n/e'P^{\perp}e)b_n + a_n(e+\delta_o)'P_1^{\perp}(e+\delta_o)/n \;.$$

By Theorem 1, $n\ a_n b_n$ and $n(n/e'P^{\perp}e)b_n$ converge in probability to zero. Now $a_n(e+\delta_o)'(e+\delta_o)/n$ bounds the last term from above as $P_1^{\perp}$

is idempotent. As in the proof of Lemma 2, $(e+\delta_o)'(e+\delta_o)/n$ converges almost surely to σ^2 and $n\, a_n$ converges almost surely to zero by Theorem 1.

Set $z = (1/\sigma)e$, $Y = (1/\sigma)\delta_o$, and $R = P-P_1$. The random variables $(z_1, z_2, \ldots, z_n)$ are independent with density $n(t; 0, 1)$. For an arbitrary constant b, the random variable $(z + bY)'R(z + bY)$ is a non-central Chi-squared with q degrees freedom and non-centrality $b^2Y'R\,Y/2$ since R is idempotent with rank q. Similarly, $(z + bY)'P^{\perp}(z + bY)$ is a non-central Chi-squared with $n-p$ degrees freedom and non-centrality $b^2Y'P^{\perp}Y/2$. These two random variables are independent because $RP^{\perp} = 0$; see Graybill [5, p. 79 ff].

Let $a > 0$.

$$P[X > a + 1] = P[(z + Y)'P_1^{\perp}(z + Y) > (a + 1)z'P^{\perp}z]$$

$$= P[(z + Y)'R(z + Y) > a\, z'P^{\perp}z - 2Y'P^{\perp}z - Y'P^{\perp}Y]$$

$$= P[(z + Y)'R(z + Y) > a(z - a^{-1}Y)'P^{\perp}(z - a^{-1}Y) - (1+a^{-1})Y'P^{\perp}Y]$$

$$= \int_0^{\infty} P[t > a(z - a^{-1}Y)'P^{\perp}(z - a^{-1}Y) - (1 + a^{-1})Y'P^{\perp}Y]$$

$$\times\, g(t; q, Y'R\,Y/2)dt$$

$$= \int_0^{\infty} P[(z - a^{-1}Y)'P^{\perp}(z - a^{-1}Y) < (t + (1 + a^{-1})Y'P^{\perp}Y)/a]$$

$$\times\, g(t; q, Y'R\,Y/2)dt$$

$$= \int_0^{\infty} G(t/a + (a+1)Y'P^{\perp}Y/a^2; n-p, Y'P^{\perp}Y/[2\, a^2])$$

$$\times\, g(t; q, Y'R\,Y/2)dt\ .$$

By substituting $x = a-1$, $\lambda_1 = Y'R\,Y/2$, and $\lambda_2 = Y'P^{\perp}Y/2$ one obtains the form of the distribution function for $x > 1$.

The derivations for the remaining cases are analogous and are omitted. □

Proof of Theorem 3. The random variable S may be written

$$S = \frac{n-p}{nq}(\hat{\tau}-\tau_o)'C_{22}^{-1}(\hat{\tau}-\tau_o)/\hat{\sigma}^2 + \frac{n-p}{nq}(\hat{\tau}-\tau_o)'(\hat{C}_{22}^{-1} - C_{22}^{-1})(\hat{\tau}-\tau_o)/\hat{\sigma}^2 .$$

The second term on the right will be denoted by u_n; u_n converges in probability to zero because $\sqrt{n}(\hat{\tau} - \tau^*)$ is bounded in probability by Theorem 1 and Lemma 3, $\sqrt{n}(\tau_o - \tau^*)$ converges to a finite limit by assumption, $\hat{\sigma}^2$ converges almost surely to σ^2, and $n^{-1}(\hat{C}_{22}^{-1} - C_{22}^{-1})$ converges almost surely to the zero matrix by Lemma 1, iii.

Using the expression for $\hat{\theta}$ given in Theorem 1

$$S = \frac{n-p}{nq} Z_2'C_{22}^{-1}Z_2/\hat{\sigma}^2 + \frac{n-p}{nq}[2\,\alpha_2'\,C_{22}^{-1}Z_2 + \alpha_2'C_{22}^{-1}\alpha_2]/\hat{\sigma}^2 + u_n$$

where Z_2 and α_2 are formed from $(F'F)^{-1}F'e + \theta^* - \theta_o$ and α_n by deleting the first r rows. Denote the second term on the right by v_n. As noted previously $\sqrt{n}\,Z_2$ is bounded in probability, $\sqrt{n}\,\alpha_2$ converges in probability to the zero vector as stated in Theorem 1, and by Lemma 1, i the matrix $n^{-1}C_{22}^{-1}$ converges. Consequently v_n converges in probability to zero.

We now use the expression for $1/\hat{\sigma}^2$ given in Theorem 1 to obtain

$$S = \frac{Z_2'C_{22}^{-1}Z_2/q}{e'P^{\perp}e/(n-p)} + \frac{(n-p)}{nq} Z_2'C_{22}^{-1}Z_2 a_n + v_n + u_n$$

where a_n is as in Theorem 1. For the reasons noted previously, $Z_2'C_{22}^{-1}Z_2$ is bounded in probability while a_n converges almost surely to zero. Thus $S = Y + Y_n$ where Y_n converges in probability to zero and Y has the familiar form of the test statistic used in linear regression [5, pp. 77-78]. $\square$