NONLINEAR STATISTICAL MODELS

by

A. Ronald Gallant

CHAPTER 1. Univariate Nonlinear Regression

This printing is being circulated for discussion. Please send comments and report errors to the following address.

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

Phone: 1-919-737-2531

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

NONLINEAR STATISTICAL MODELS

Table of Contents

		Anticipated Completion Date
1.	Univariate Nonlinear Regression	Completed
	1.0 Preface	
	1.1 Introduction	
	1.2 Taylor's Theorem and Matters of Notation	
	1.3 Statistical Properties of Least Squares Estimators	
	1.4 Methods of Computing Least Squares Estimators	
	1.5 Hypothesis Testing	
	1.6 Confidence Intervals	
	1.7 References	
	1.8 Index	
2.	Univariate Nonlinear Regression: Special Situations	December 1985
3.	A Unified Asymptotic Theory of Nonlinear Statistical Models	Completed
	3.0 Preface	
	3.1 Introduction	
	3.2 The Data Generating Model and Limits of Cesaro Sums	
	3.3 Least Mean Distance Estimators	
	3.4 Method of Moments Estimators	
	3.5 Tests of Hypotheses	
	3.6 Alternative Representations of a Hypothesis	
	3.7 Random Regressors	
	3.8 Constrained Estimation	
	3.9 References	
	3.10 Index	
4.	Univariate Nonlinear Regression: Asymptotic Theory	June 1983
5.	Multivariate Linear Models: Review	December 1983
6.	Multivariate Nonlinear Models	December 1983
7.	Linear Simultaneous Equations Models: Review	June 1984
8.	Nonlinear Simultaneous Equations Models	June 1984

CHAPTER 1. Univariate Nonlinear Regression

The nonlinear regression model with a univariate dependent variable is more frequently used in applications than any of the other methods discussed in this book. Moreover, these other methods are for the most part fairly straightforward extensions of the ideas of univariate nonlinear regression. Accordingly, we shall take up this topic first and consider it in some detail.

In this chapter, we shall present the theory and methods of univariate nonlinear regression by relying on analogy with the theory and methods of linear regression, on examples, and on Monte-Carlo illustrations. The formal mathematical verifications are presented in subsequent chapters. The topic lends itself to this treatment as the role of the theory is to justify some intuitively obvious linear approximations derived from Taylor's expansions. Thus one can get the main ideas across first and save the theoretical details until later. This is not to say that the theory is unimportant. Intuition is not entirely reliable and some surprises are uncovered by careful attention to regularity conditions and mathematical detail.

As a practical matter, the computations for nonlinear regression methods must be performed using either a scientific subroutine library such as IMSL or NAg Libraries or a statistical package with nonlinear capabilities such as SAS, BMDP, TROLL, or TSP. Hand calculator computations are out of the question. One who writes his own code with repetitive use in mind will probably produce something similar to the routines found in a scientific subroutine library. Thus, a scientific subroutine library or a statistical package are effectively the two practical alternatives. Granted that scientific subroutine packages are far more flexible than statistical packages and are usually nearer to the state of the art of numerical analysis than the

statistical packages, they nonetheless make poor pedigological devices. The illustrations would consist of lengthy FORTRAN codes with the main line of thought obscured by bookkeeping details. For this reason we have chosen to illustrate the computations with a statistical package, namely SAS.

1. INTRODUCTION

One of the most common situations in statistical analysis is that of data which consist of observed, univariate responses y_t known to be dependent on corresponding k-dimensional inputs x_t. This situation may be represented by the regression equations

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

where $f(x,\theta)$ is the known response function, θ^o is a p-dimensional vector of unknown parameters, and the e_t represent unobservable observational or experimental errors. We write θ^o to emphasize that it is the true, but unknown, value of the parameter vector θ that is meant; θ itself is used to denote instances when the parameter vector is treated as a variable as, for instance, in differentiation. The errors are assumed to be independently and identically distributed with mean zero and unknown variance σ^2. The sequence of independent variables $\{x_t\}$ is treated as a fixed known sequence of constants, not random variables. If some components of the independent vectors were generated by a random process, then the analysis is conditional on that realization $\{x_t\}$ which obtained for the data at hand. See Section 2 of the next chapter for additional details on this point and Section 7 of the next chapter in which is displayed a device that allows one to consider the random regressor set-up as a special case in a fixed regressor theory.

Frequently, the effect of the independent variable x_t on the dependent variable y_t is adequately approximated by a response function which is linear in the parameters

$$f(x,\theta) = x'\theta = \sum_{i=1}^{p} x_i\theta_i$$

By exploiting various transformations of the independent and dependent variables, viz.

$$\phi_0(y_t) = \sum_{i=1}^{p} \phi_i(x_{it})\theta_i + e_t$$

the scope of models that are linear in the parameters can be extended considerably. But there is a limit to what can be adequately approximated by a linear model. At times a plot of the data or other data analytic considerations will indicate that a model which is not linear in its parameters will better represent the data. More frequently, nonlinear models arise in instances where a specific scientific discipline specifies the form that the data ought to follow and this form is nonlinear. For example, a response function which arises from the solution of a differential equation might assume the form

$$f(x,\theta) = \theta_1 + \theta_2 e^{x\theta_3}.$$

Another example is a set of responses that is known to be periodic in time but with an unknown period. A response function for such data is

$$f(t,\theta) = \theta_1 + \theta_2\cos(\theta_4 t) + \theta_3\sin(\theta_4 t).$$

A univariate linear regression model, for our purposes, is a model that can be put in the form

$$\phi_0(y_t) = \sum_{i=1}^{p} \phi_i(x_t)\theta_i + e_t.$$

A univariate nonlinear regression model is of the form

$$\phi_0(y_t) = f(x_t,\theta) + e_t$$

but since the transformation ϕ_0 can be absorbed into the definition of the dependent variable, the model

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

is sufficiently general. Under these definitions a linear model is a special case of the nonlinear model in the same sense that a central chi-square distribution is a special case of the non-central chi-square distribution. This is somewhat of an abuse of language as one ought to say regression model and linear regression model rather than nonlinear regression model and (linear) regression model to refer to these two categories. But this usage is long established and it is senseless to seek change now.

EXAMPLE 1. The example that we shall use most frequently in illustration has the response function

$$f(x,\theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}.$$

The vector valued input or independent variable is

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

and the vector valued parameter is

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix}$$

so that for this response function k = 3 and p = 4. A set of observed responses and inputs for this model which will be used to illustrate the computations is given in Table 1. The inputs correspond to a one-way "treatment-control" design that uses experimental material whose age ($=x_3$) affects the response exponentially. That is, the first observation

$$x_1 = (0,1,6.28)$$

represents experimental material with attained age $x_3 = 6.28$ months that was (randomly) allocated to the control group and has expected response.

$$f(x_1,\theta^o) = \theta_2^o + \theta_4^o\, e^{6.28\theta_3^o}$$

Similarly, the second observation

$$x_2 = (1,1,9.86)$$

represents an allocation of material with attained age $x_3 = 9.86$ to the treatment group; with expected response

$$f(x_2,\theta^o) = \theta_1^o + \theta_2^o + \theta_4^o\, e^{9.86\theta_3^o};$$

Table 1. Data Values for Example 1.

t	Y	X1	X2	X3
1	0.98610	1	1	6.28
2	1.03848	0	1	9.86
3	0.95482	1	1	9.11
4	1.04184	0	1	8.43
5	1.02324	1	1	8.11
6	0.90475	0	1	1.82
7	0.96263	1	1	6.58
8	1.05026	0	1	5.02
9	0.98861	1	1	6.52
10	1.03437	0	1	3.75
11	0.98982	1	1	9.86
12	1.01214	0	1	7.31
13	0.66768	1	1	0.47
14	0.55107	0	1	0.07
15	0.96822	1	1	4.07
16	0.98823	0	1	4.61
17	0.59759	1	1	0.17
18	0.99418	0	1	6.99
19	1.01962	1	1	4.39
20	0.69163	0	1	0.39
21	1.04255	1	1	4.73
22	1.04343	0	1	9.42
23	0.97526	1	1	8.90
24	1.04969	0	1	3.02
25	0.80219	1	1	0.77
26	1.01046	0	1	3.31
27	0.95196	1	1	4.51
28	0.97658	0	1	2.65
29	0.50811	1	1	0.08
30	0.91840	0	1	6.11

and so on. The parameter θ_1^o is, then, the treatment effect. The data of Table 1 are simulated. ∎

EXAMPLE 2. Quite often, nonlinear models arise as solutions of a system of differential equations. The following linear system has been used so often in the nonlinear regression literature (Box and Lucus (1959), Guttman and Meeter (1964), Gallant (1980)) that it might be called the standard pedagogical example.

Linear System

$$(d/dx)A(x) = -\theta_1 A(x)$$

$$(d/dx)B(x) = \theta_1 A(x) - \theta_2 B(x)$$

$$(d/dx)C(x) = \theta_2 B(x)$$

Boundary Conditions

$$A(x) = 1,\ B(x) = C(x) = 0 \text{ at time } x = 0$$

Parameter Space

$$\theta_1 > \theta_2 > 0$$

Solution, $\theta_1 > \theta_2$

$$A(x) = e^{-\theta_1 x}$$

$$B(x) = (\theta_1 - \theta_2)^{-1}(\theta_1 e^{-\theta_2 x} - \theta_1 e^{-\theta_1 x})$$

$$C(x) = 1 - (\theta_1 - \theta_2)^{-1}(\theta_1 e^{-\theta_2 x} - \theta_2 e^{-\theta_1 x})$$

Solution, $\theta_1 = \theta_2$

$$A(x) = e^{-\theta_1 x}$$

$$B(x) = \theta_1 x e^{-\theta_1 x}$$

$$C(x) = 1 - e^{-\theta_1 x} - \theta_1 x e^{-\theta_1 x}$$

Systems such as this arise in compartment analysis where the rate of flow of a substance from compartment A into compartment B is a constant proportion θ_1 of the amount A(x) present in compartment A at time x. Similarly, the rate of flow from B to C is a constant proportion θ_2 of the amount B(x) present in compartment B at time x. The rate of change of the quantities within each compartment is described by the system of linear differential equations. In chemical kinetics, this model describes a reaction where substance A decomposes at a reaction rate of θ_1 to form substance B which in turn decomposes at a rate θ_2 to form substance C. There are a great number of other instances where linear systems of differential equations such as this arise.

Following Guttman and Meeter (1964) we shall use the solutions for B(x) and C(x) to construct two nonlinear models which they assert "represent fairly well the extremes of near linearity and extreme nonlinearity." These two models are set forth immediately below. The design points and parameter settings are those of Guttman and Meeter (1964).

Model B

$$f(x,\theta) = \begin{cases} \theta_1(e^{-x\theta_2} - e^{-x\theta_1})/(\theta_1 - \theta_2) & \theta_1 \neq \theta_2 \\ \theta_1 \times e^{-x\theta_1} & \theta_1 = \theta_2 \end{cases}$$

$$\theta^o = (1.4,\ .4)$$

$$\{x_t\} = \{.25,\ .5,\ 1,\ 1.5,\ 2,\ 4,\ .25,\ .5,\ 1,\ 1.5,\ 2,\ 4\}$$

$$n = 12$$

$$\sigma^2 = (.025)^2$$

Model C

$$f(x,\theta) = \begin{cases} 1 - (\theta_1 e^{-x\theta_2} - \theta_2 e^{-x\theta_1})/(\theta_1 - \theta_2) & \theta_1 \neq \theta_2 \\ 1 - e^{-x\theta_1} - x\theta_1 e^{-x\theta_1} & \theta_1 = \theta_2 \end{cases}$$

$$\theta^o = (1.4,\ .4)$$

$$\{x_t\} = \{1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6\}$$

$$n = 12$$

$$\sigma^2 = (.025)^2$$

Table 2. Data Values for Example 2.

t	Y	X
	Model B	
1	0.316122	0.25
2	0.421297	0.50
3	0.601996	1.00
4	0.573076	1.50
5	0.545661	2.00
6	0.281509	4.00
7	0.273234	0.25
8	0.415292	0.50
9	0.603644	1.00
10	0.621614	1.50
11	0.515790	2.00
12	0.278507	4.00
	Model C	
1	0.137790	1
2	0.409262	2
3	0.639014	3
4	0.736366	4
5	0.786320	5
6	0.893237	6
7	0.163208	1
8	0.372145	2
9	0.599155	3
10	0.749201	4
11	0.835155	5
12	0.905845	6

2. TAYLOR'S THEOREM AND MATTERS OF NOTATION

In what follows, a matrix notation for certain concepts in differential calculus leads to a more compact and readable exposition. Suppose that $s(\theta)$ is a real valued function of a p-dimensional argument θ. The notation $(\partial/\partial\theta)s(\theta)$ denotes the gradient of $s(\theta)$,

$$(\partial/\partial\theta)s(\theta) = \underset{p}{\begin{pmatrix} (\partial/\partial\theta_1)s(\theta) \\ (\partial/\partial\theta_2)s(\theta) \\ \vdots \\ (\partial/\partial\theta_p)s(\theta) \end{pmatrix}} \quad \underset{1}{}$$

a p by 1 (column) vector with typical element $(\partial/\partial\theta_i)s(\theta)$. Its transpose is denoted by

$$(\partial/\partial\theta')s(\theta) = \underset{1}{[(\partial/\partial\theta_1)s(\theta),\ (\partial/\partial\theta_2)s(\theta),\ \ldots,\ (\partial/\partial\theta_p)s(\theta)]} \quad \underset{p}{}.$$

Suppose that all second order derivatives of $s(\theta)$ exist. They can be arranged in a p by p matrix, known as the Hessian matrix of the function $s(\theta)$,

$$(\partial^2/\partial\theta\partial\theta')s(\theta) = \underset{p}{\begin{pmatrix} (\partial^2/\partial\theta_1^2)s(\theta) & (\partial^2/\partial\theta_1\partial\theta_2)s(\theta) & \ldots & (\partial^2/\partial\theta_1\partial\theta_p)s(\theta) \\ (\partial^2/\partial\theta_2\partial\theta_1)s(\theta) & (\partial^2/\partial\theta_2^2)s(\theta) & \ldots & (\partial^2/\partial\theta_2\partial\theta_p)s(\theta) \\ \vdots & & & \\ (\partial^2/\partial\theta_p\partial\theta_1)s(\theta) & (\partial^2/\partial\theta_p\partial\theta_2)s(\theta) & \ldots & (\partial^2/\partial^2\theta_p)s(\theta) \end{pmatrix}} \quad \underset{p}{}$$

If the second order derivatives of $s(\theta)$ are continuous functions in θ then the Hessian matrix is symmetric (Young's theorem).

Let $f(\theta)$ be an n by 1 (column) vector-valued function of a p-dimensional argument θ. The Jacobian of

$$f(\theta) = \underset{n \quad 1}{\begin{pmatrix} f_1(\theta) \\ f_2(\theta) \\ \vdots \\ f_n(\theta) \end{pmatrix}}$$

is the n by p matrix

$$(\partial/\partial\theta')f(\theta) = \underset{n}{\begin{pmatrix} (\partial/\partial\theta_1)f_1(\theta) & (\partial/\partial\theta_2)f_1(\theta) & \cdots & (\partial/\partial\theta_p)f_1(\theta) \\ (\partial/\partial\theta_1)f_2(\theta) & (\partial/\partial\theta_2)f_2(\theta) & \cdots & (\partial/\partial\theta_p)f_2(\theta) \\ \vdots & & & \\ (\partial/\partial\theta_1)f_n(\theta) & (\partial/\partial\theta_2)f_n(\theta) & \cdots & (\partial/\partial\theta_p)f_n(\theta) \end{pmatrix}}_{p}$$

Let $h'(\theta)$ be an n by 1 (row) vector-valued function

$$h'(\theta) = [h_1(\theta), h_2(\theta), \ldots, h_n(\theta)].$$

Then

$$(\partial/\partial\theta)h'(\theta) = \underset{p}{\begin{pmatrix} (\partial/\partial\theta_1)h_1(\theta) & (\partial/\partial\theta_1)h_2(\theta) & \cdots & (\partial/\partial\theta_1)h_n(\theta) \\ (\partial/\partial\theta_2)h_1(\theta) & (\partial/\partial\theta_2)h_2(\theta) & & (\partial/\partial\theta_2)h_n(\theta) \\ \vdots & \vdots & & \vdots \\ (\partial/\partial\theta_p)h_1(\theta) & (\partial/\partial\theta_p)h_2(\theta) & & (\partial/\partial\theta_p)h_n(\theta) \end{pmatrix}}_{n}$$

In this notation, the following rule governs matrix transposition:

$$[(\partial/\partial\theta')f(\theta)]' = (\partial/\partial\theta)f'(\theta)$$

And the Hessian matrix of $s(\theta)$ can be obtained by successive differentiation variously as:

$$(\partial^2/\partial\theta\partial\theta')s(\theta) = (\partial/\partial\theta)[(\partial/\partial\theta')s(\theta)]$$

$$= (\partial/\partial\theta)[(\partial/\partial\theta)s(\theta)]'$$

$$= (\partial/\partial\theta')[(\partial/\partial\theta)s(\theta)] \quad \text{(if symmetric)}$$

$$= (\partial/\partial\theta')[(\partial/\partial\theta')s(\theta)]' \quad \text{(if symmetric)}$$

One has a chain rule and a composite function rule. They read as follows. If $f(\theta)$ and $h'(\theta)$ are as above then (Problem 1)

$$\underset{l \quad n \quad p}{(\partial/\partial\theta')h'(\theta)f(\theta)} = \underset{p \quad l \quad n \quad p}{h'(\theta)[(\partial/\partial\theta')f(\theta)] + f'(\theta)[(\partial/\partial\theta')h(\theta)]}$$

Let $g(\rho)$ be a p by 1 (column) vector-valued function of a r-dimensional argument ρ and let $f(\theta)$ as above: Then (Problem 2)

$$(\partial/\partial\rho')f[g(\rho)] = \underset{n}{(\partial/\partial\theta')f(\theta)\Big|_{\theta=g(\rho)}} \underset{p}{} \underset{r}{(\partial/\partial\rho')g(\rho)}$$

The set of nonlinear regression equations

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

may be written in a convenient vector form

$$y = f(\theta^o) + e$$

by adopting conventions analogous to those employed in linear regression; namely

$$y = \underset{n \quad 1}{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}},$$

$$f(\theta) = \underset{n}{\begin{pmatrix} f(x_1,\theta) \\ f(x_2,\theta) \\ \vdots \\ f(x_n,\theta) \end{pmatrix}}_{1},$$

$$e = \underset{n}{\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}}_{1}.$$

The sum of squared deviations

$$SSE(\theta) = \sum_{t=1}^{n} [y_t - f(x_t,\theta)]^2$$

of the observed y_t from the predicted value $f(x_t,\theta)$ corresponding to a trial value of the parameter θ becomes

$$SSE(\theta) = [y - f(\theta)]'[y - f(\theta)] = \|y - f(\theta)\|^2$$

in this vector notation.

The estimators employed in nonlinear regression can be characterized as linear and quadratic forms in the vector e which are similar in appearance to those that appear in linear regression to within an error of approximation that becomes negligible in large samples. Let

$$F(\theta) = (\partial/\partial\theta')f(\theta);$$

that is, $F(\theta)$ is the matrix with typical element $(\partial/\partial\theta_j)f(x_t,\theta)$ where t is the row index and j is the column index. The matrix $F(\theta^o)$ plays the same role in these linear and quadratic forms as the design matrix X in the linear regression.

$$z = X\beta + e.$$

The appropriate analogy is obtained by setting $z = y - f(\theta^o) + F(\theta^o)\theta^o$ and setting $X = F(\theta^o)$. Malinvaud (1970, Ch. 9) terms this equation the "linear pseudo-model." For simplicity we shall write F for the matrix $F(\theta)$ when it is evaluated at $\theta=\theta^o$;

$$F \equiv F(\theta^o).$$

Let us illustrate these notations with Example 1.

EXAMPLE 1 (continued). Direct application of the definitions of y and $f(\theta)$ yields

$$y = \begin{pmatrix} 0.98610 \\ 1.03848 \\ 0.95482 \\ 1.04184 \\ \vdots \\ 0.50811 \\ 0.91840 \end{pmatrix},$$

$$f(\theta) = \begin{pmatrix} \theta_1 + \theta_2 + \theta_4 e^{6.28\theta_3} \\ \theta_2 + \theta_4 e^{9.86\theta_3} \\ \theta_1 + \theta_2 + \theta_4 e^{9.11\theta_3} \\ \theta_2 + \theta_4 e^{8.43\theta_3} \\ \vdots \\ \theta_1 + \theta_2 + \theta_4 e^{0.08\theta_3} \\ \theta_2 + \theta_4 e^{6.11\theta_3} \end{pmatrix}, \qquad \begin{matrix} 30 \\ \\ \\ \\ \\ \\ \\ \end{matrix} \quad \begin{matrix} 1 \\ \\ \\ \\ \\ \\ \\ \end{matrix}$$

Since

$$(\partial/\partial\theta_1)f(x,\theta) = (\partial/\partial\theta_1)(\theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}) = x_1$$

$$(\partial/\partial\theta_2)f(x,\theta) = (\partial/\partial\theta_2)(\theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}) = x_2$$

$$(\partial/\partial\theta_3)f(x,\theta) = (\partial/\partial\theta_3)(\theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}) = \theta_4 x_3 e^{\theta_3 x_3}$$

$$(\partial/\partial\theta_4)f(x,\theta) = (\partial/\partial\theta_4)(\theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}) = e^{\theta_3 x_3}$$

The Jacobian of $f(\theta)$ is

$$F(\theta) = \begin{pmatrix} 1 & 1 & \theta_4(6.28)e^{6.28\theta_3} & e^{6.28\theta_3} \\ 0 & 1 & \theta_4(9.86)e^{9.86\theta_3} & e^{9.86\theta_3} \\ 1 & 1 & \theta_4(9.11)e^{9.11\theta_3} & e^{9.11\theta_3} \\ 0 & 1 & \theta_4(8.43)e^{8.43\theta_3} & e^{8.43\theta_3} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \theta_4(0.08)e^{0.08\theta_3} & e^{0.08\theta_3} \\ 0 & 1 & \theta_4(6.11)e^{6.11\theta_3} & e^{6.11\theta_3} \end{pmatrix}$$

30 4

Taylor's theorem, as we shall use it, reads as follows:

Taylor's Theorem: Let $s(\theta)$ be a real valued function defined over Θ. Let Θ be an open, convex subset of R^p; R^p denotes p-dimensional Euclidean space. Let θ^o be some point in Θ.

If $s(\theta)$ is once continuously differentiable on Θ then

$$s(\theta) = s(\theta^o) + \sum_{i=1}^{p} [(\partial/\partial\theta_i)s(\bar{\theta})](\theta_i - \theta_i^o)$$

or, in vector notation,

$$s(\theta) = s(\theta^o) + [(\partial/\partial\theta)s(\bar{\theta})]'(\theta - \theta^o)$$

for some $\bar{\theta} = \lambda\theta^o + (1-\lambda)\theta$ where $0 \leq \lambda \leq 1$.

If $s(\theta)$ is twice continuously differentiable on Θ then

$$s(\theta) = s(\theta^o) + \sum_{i=1}^{p} [(\partial/\partial\theta_i)s(\theta^o)](\theta_i - \theta_i^o) + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} (\theta_i - \theta_i^o)[(\partial^2/\partial\theta_i\partial\theta_j)s(\bar{\theta})](\theta_j - \theta_j^o)$$

or, in vector notation,

$$s(\theta) = s(\theta^o) + [(\partial/\partial\theta)s(\theta^o)]'(\theta - \theta^o) + \frac{1}{2}(\theta - \theta^o)'[(\partial/\partial\theta\partial\theta')s(\bar{\theta})](\theta - \theta^o)$$

for some $\bar{\theta} = \lambda\theta^o + (1-\lambda)\theta$ where $0 \leq \lambda \leq 1$. ∎

Applying Taylor's theorem to $f(x_t,\theta)$ we have

$$f(x,\theta) = f(x,\theta^o) + [(\partial/\partial\theta)f(x,\theta^o)]'(\theta - \theta^o) + \frac{1}{2}(\theta - \theta^o)'[(\partial^2/\partial\theta\partial\theta')f(x,\bar{\theta})](\theta - \theta^o)$$

implicitly assuming that $f(x,\theta)$ is twice continuously differentiable on some open, convex set Θ. Note that $\bar{\theta}$ is a function of both x and θ, $\bar{\theta} = \bar{\theta}(x,\theta)$. Applying this formula row by row to the vector $f(\theta)$ we have the approximation

$$f(\theta) = f(\theta^o) + [(\partial/\partial\theta')f(\theta^o)](\theta - \theta^o) + R(\theta - \theta^o)$$

where a typical row of R is

$$r_t' = \frac{1}{2}(\theta - \theta^o)'[(\partial^2/\partial\theta\partial\theta')f(x_t,\bar{\theta})]\Big|_{\bar{\theta} = \bar{\theta}(x_t,\theta)};$$

alternatively

$$f(\theta) = f(\theta^o) + F(\theta^o)(\theta - \theta^o) + R(\theta - \theta^o).$$

Using the previous formulas,

$$(\partial/\partial\theta')SSE(\theta) = (\partial/\partial\theta')[y - f(\theta)]'[y - f(\theta)]$$

$$= [y - f(\theta)]'(\partial/\partial\theta')[y - f(\theta)] + [y - f(\theta)]'(\partial/\partial\theta')[y - f(\theta)]$$

$$= 2[y - f(\theta)]'[-(\partial/\partial\theta')f(\theta)]$$

$$= -2[y - f(\theta)]'F(\theta)$$

The least squares estimator is that value $\hat{\theta}$ that minimizes $SSE(\theta)$ over the parameter space Θ. If $SSE(\theta)$ is once continuously differentiable on some open set Θ^o with $\theta \in \Theta^o \subset \Theta$, then $\hat{\theta}$ satisfies the "normal equations"

$$F'(\hat{\theta})[y - f(\hat{\theta})] = 0.$$

This is because $(\partial/\partial\theta)SSE(\hat{\theta}) = 0$ at any local optimum. In linear regression,

$$z = X\beta + e,$$

least square residuals $\hat{e}$ computed as

$$\hat{e} = y - X\hat{\beta}, \qquad \hat{\beta} = (X'X)^{-1}X'y$$

are orthogonal to the columns of X, viz.,

$$X'\hat{e} = 0.$$

In nonlinear regression, least squares residuals are orthogonal to the columns of the Jacobian of $f(\theta)$ evaluated at $\theta = \hat{\theta}$, viz.,

$$F'(\hat{\theta})[y - f(\hat{\theta})] = 0.$$

PROBLEMS

1. (Chain rule). Show that

$$(\partial/\partial\theta')h'(\theta)f(\theta) = h'(\theta)(\partial/\partial\theta')f(\theta) + f'(\theta)(\partial/\partial\theta')h(\theta)$$

by computing $(\partial/\partial\theta_i)\sum_{k=1}^{n} h_k(\theta)f_k(\theta)$ by the chain rule for $i=1,2,\ldots,p$ to obtain

$$(\partial/\partial\theta')h'(\theta)f(\theta) = \sum_{k=1}^{n} h_k(\theta)(\partial/\partial\theta')f_k(\theta) + \sum_{k=1}^{n} f_k(\theta)(\partial/\partial\theta')h_k(\theta)$$

Note that $(\partial/\partial\theta')f_k(\theta)$ is the k-th row of $(\partial/\partial\theta')f(\theta)$.

2. (Composite function rule). Show that

$$(\partial/\partial\rho')f[g(\rho)] = \{(\partial/\partial\theta')f[g(\rho)]\}(\partial/\partial\rho')g(\rho)$$

by computing the (i,j) element of $(\partial/\partial\rho')f[g(\rho)]$, $(\partial/\partial\rho_j)f_i[g(\rho)]$ and then applying the definition of matrix multiplication.

3. STATISTICAL PROPERTIES OF LEAST SQUARES ESTIMATORS

The least squares estimator of the unknown parameter θ^o in the nonlinear model

$$y = f(\theta^o) + e$$

is the p by 1 vector $\hat{\theta}$ that minimizes

$$SSE(\theta) = [y - f(\theta)]'[y - f(\theta)] = \|y - f(\theta)\|^2.$$

The estimate of the variance of the errors e_t corresponding to the least squares estimator $\hat{\theta}$ is

$$s^2 = SSE(\hat{\theta})/(n - p).$$

In Chapter 4 we shall show that

$$\hat{\theta} = \theta^o + (F'F)^{-1}F'e + o_p(1/\sqrt{n})$$

$$s^2 = e'[I - F(F'F)^{-1}F']e/(n-p) + o_p(1/n)$$

where, recall, $F = F(\theta^o) = (\partial/\partial\theta')f(\theta^o)$ = matrix with typical element $(\partial/\partial\theta')f(x_t,\theta^o)$. The notation $o_p(a_n)$ denotes a (possibly) matrix-valued random variable $X_n = o_p(a_n)$ with the property that each element X_{ijn} satisfies

$$\lim_{n\to\infty} P[|X_{ijn}/a_n| > \varepsilon] = 0$$

for any $\varepsilon > 0$; $\{a_n\}$ is some sequence of real numbers, the most frequent choices being $a_n \equiv 1$, $a_n = 1/\sqrt{n}$, and $a_n = 1/n$.

These equations suggest that a good approximation to the joint distribution of $(\hat{\theta},s^2)$ can be obtained by simply ignoring the terms $o_p(1/\sqrt{n})$ and $o_p(1/n)$. Then by noting the similarity of the equations

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

$$s^2 = e'[I - F(F'F)^{-1}F']e/(n - p)$$

with the equations that arise in linear models theory and assuming normal errors we have approximately that $\hat{\theta}$ has the p-dimensional multivariate normal distribution with mean θ^o and variance-covariance matrix $\sigma^2(F'F)^{-1}$;

$$\hat{\theta} \sim N_p[\theta^o,\sigma^2(F'F)^{-1}];$$

$(n-p)s^2/\sigma^2$ has the chi-squared distribution with $(n-p)$ degrees of freedom,

$$(n-p)s^2/\sigma^2 \sim \chi^2(n-p);$$

and s^2 and $\hat{\theta}$ are independent so that the joint distribution of $(\hat{\theta},s^2)$ is the product of the marginal distributions. In applications, $(F'F)^{-1}$ must be approximated by the matrix

$$\hat{C} = [F'(\hat{\theta})F(\hat{\theta})]^{-1}.$$

The alternative to this method of obtaining an approximation to the distribution of $\hat{\theta}$--characterization coupled with a normality assumption--is to use conventional asymptotic arguments. One finds that $\hat{\theta}$ converges almost surely to θ^o, s^2 converges almost surely to σ^2, $(1/n)F'(\hat{\theta})F(\hat{\theta})$ converges almost surely to a matrix Ω, and that $\sqrt{n}(\hat{\theta} - \theta^o)$ is asymptotically normally distributed as the p-variate normal with mean zero and variance-covariance matrix $\sigma^2\Omega^{-1}$,

$$\sqrt{n}(\hat{\theta} - \theta^o) \xrightarrow{L} N_p(0,\sigma^2\Omega^{-1}).$$

The normality assumption is not needed. Let

$$\hat{\Omega} = (1/n)F'(\hat{\theta})F(\hat{\theta}).$$

Following the characterization/normality approach it is natural to write

$$\hat{\theta} \dot{\sim} N_p(\theta^o, s^2\hat{C}) \qquad \left(= N_p[\theta^o, s^2(1/n)\hat{\Omega}^{-1}] \right)$$

Following the asymptotic normality approach it is natural to write

$$\sqrt{n}(\hat{\theta} - \theta^o) \dot{\sim} N_p(0, s^2\hat{\Omega}^{-1}) \qquad \left(= N_p(0, s^2n\hat{C}) \right);$$

natural perhaps even to drop the degrees of freedom correction and use

$$\hat{\sigma}^2 = (1/n)SSE(\hat{\theta})$$

to estimate σ^2 instead of s^2. The practical difficulty with this is that one can never be sure of the scaling factors in computer output. Natural combinations to report are:

$$\hat{\theta},\ s^2,\ \hat{C};$$
$$\hat{\theta},\ s^2,\ s^2\hat{C};$$
$$\hat{\theta},\ \hat{\sigma}^2,\ \hat{\Omega}^{-1};$$
$$\hat{\theta},\ \hat{\sigma}^2,\ \hat{\sigma}^2\ \hat{\Omega}^{-1};$$

and so on. The documentation usually leaves some doubt in the reader's mind as to what is actually printed. Probably, the best strategy is to run the program using Example 1 and resolve the issue by comparison with the results reported in the next section.

As in linear regression, the practical importance of these distributional properties is their use to set confidence intervals on the unknown parameters θ_i^o $(i=1,2,\ldots,p)$ and to test hypotheses. For example, a 95% confidence interval may be found for θ_i^o from the .025 critical value $t_{.025}$ of the t-distribution with n-p degrees of freedom as

$$\hat{\theta}_i \pm t_{.025}\sqrt{s^2\hat{c}_{ii}}.$$

Similarly, the hypothesis H: $\theta_i^o = \theta_i^*$ may be tested against the alternative A: $\theta_i^o \neq \theta_i^*$ at the 5% level of significance by comparing

$$|\tilde{t}_i| = |\hat{\theta}_i - \theta_i^*|/\sqrt{s^2\hat{c}_{ii}}$$

with $|t_{.025}|$ and rejecting H when $|\tilde{t}_i| > |t_{.025}|$; $\hat{c}_{ii}$ denotes the i-th diagonal element of the matrix $\hat{C}$. The next few paragraphs are an attempt to convey an intuitive feel for the nature of the regularity conditions used to obtain these results; the reader is reminded once again that they are presented with complete rigor in Chapter 4.

The sequence of input vectors $\{x_t\}$ must behave properly as n tends to infinity. Proper behavior is obtained when the components x_{it} of x_t are chosen either by random sampling from some distribution or (possibly disproportionate) replication of a fixed set of points. In the latter case, some set of points $a_0, a_1, \ldots, a_{T-1}$ is chosen and the inputs assigned according to $x_{it} = a_{(t \bmod T)}$. Disproportionality is accomplished by allowing some of the a_i to be equal. More general schemes than these are permitted--see Section 2 of Chapter 3 for full details--but this is enough to gain a feel for the sort of stability that $\{x_t\}$ ought to exhibit. Consider, for instance, the data generating scheme of Example 1.

EXAMPLE 1 (continued). The first two coordinates x_{1t}, x_{2t} of $x_t = (x_{1t}, x_{2t}, x_{3t})'$ consist of replication of a fixed set of design points determined by the design structure:

$$
\begin{aligned}
(x_1,x_2)_1 &= (1,1), \\
(x_1,x_2)_2 &= (0,1), \\
&\vdots \\
(x_1,x_2)_t &= (1,1), \quad \text{if t is odd} \\
(x_1,x_2)_t &= (0,1), \quad \text{if t is even} \\
&\vdots
\end{aligned}
$$

That is,

$$(x_1,x_2)_t = a_{(t \bmod 2)}$$

with

$$a_0 = (0,1),$$
$$a_1 = (1,1)$$

The covariate x_{3t} is the age of the experimental material and is conceptually a random sample from the age distribution of the population due to the random allocation of experimental units to treatments. In the simulated data of Table 1, x_{3t} was generated by random selection from the uniform distribution on the interval [0,10]. In a practical application one would probably not know the age distribution of the experimental material but would be prepared to assume that x_3 was distributed according to a continuous distribution function that has a density $p_3(x)$ which is positive everywhere on some known interval [0,b], there being some doubt as to how much probability mass was to the right of b. I

The response function $f(x,\theta)$ must be continuous in the argument (x,θ); that is, if $\lim_{i\to\infty} (x_i,\theta_i) = (x^*,\theta^*)$ (in Euclidean norm on R^{k+p}) then $\lim_{i\to\infty} f(x_i,\theta_i) = f(x^*,\theta^*)$. The first partial derivatives $(\partial/\partial\theta_i)f(x,\theta)$ must be continuous in (x,θ) and the second partial derivatives $(\partial^2/\partial\theta_i\,\partial\theta_j)f(x,\theta)$ must be continuous in (x,θ). These smoothness requirements are due to the heavy use of Taylor's theorem in Chapter 3. Some relaxation of the second derivative requirement is possible (Gallant, 1973). Quite probably, further relaxation is possible (Huber, 1982).

There remain two further restrictions on the limiting behavior of the response function and its derivatives which roughly correspond to estimability considerations in linear models. The first is that

$$s(\theta) = \lim_{n\to\infty} (1/n) \sum_{t=1}^{n} [f(x_t,\theta) - f(x_t,\theta^o)]^2$$

has a unique minimum at $\theta = \theta^o$ and the second is that the matrix

$$\Omega = \lim_{n\to\infty} (1/n)F'(\theta^o)F(\theta^o)$$

be non-singular. We term these the Identification Condition and the Rank Qualification respectively. When random sampling is involved, Kolmogorov's Strong Law of Large Numbers is used to obtain the limit as we illustrate with Example 1, below. These two conditions are tedious to verify in applications and few would bother to do so. However, these conditions indirectly impose restrictions on the inputs x_t and parameter θ^o that are often easy to spot by inspection. Although θ^o is unknown in an estimation situation, when testing hypotheses one should check whether the null hypothesis violates these assumptions. If this happens, methods to circumvent the difficulty are given in the next chapter. For Example 1, either H: $\theta_3^o = 0$ or H: $\theta_4^o = 0$ will violate the Rank Qualification and the Identification Condition as we next show.

EXAMPLE 1 (continued). We shall first consider how the problems with H: $\theta_4^o = 0$ and H: $\theta_3^o = 0$ can be detected by inspection, next consider how limits are to be computed, and last how one verifies that $s(\theta) = \lim_{n\to\infty} (1/n) \sum_{t=1}^{n} [f(x_t,\theta) - f(x_t,\theta^o)]^2$ has a unique minimum at $\theta = \theta^o$.

Consider the case H: $\theta_3^o = 0$ leaving the case H: $\theta_4^o = 0$ to Problem 1. If $\theta_3^o = 0$ then

$$F(\theta) = \begin{pmatrix} 1 & 1 & \theta_4 x_{31} & 1 \\ 0 & 1 & \theta_4 x_{32} & 1 \\ 1 & 1 & \theta_4 x_{33} & 1 \\ 0 & 1 & \theta_4 x_{34} & 1 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & \theta_4 x_{3n-1} & 1 \\ 0 & 1 & \theta_4 x_{3n} & 1 \end{pmatrix}$$

$F(\theta)$ has two columns of ones and is, thus, singular. Now this fact can be noted at sight in applications; there is no need for any analysis. It is this kind of easily checked violation of the regularity conditons that one should guard against. Let us verify that the singularity carries over to the limit. Let

$$\Omega_n(\theta) = (1/n)F'(\theta)F(\theta) = (1/n)\sum_{t=1}^{n}[(\partial/\partial\theta)f(x_t,\theta)][(\partial/\partial\theta)f(x_t,\theta)]'$$

The regularity conditions of Chapter 4 guarantee that $\lim_{n\to\infty} \Omega_n(\theta)$ exists and we shall show it directly below. Put $\lambda' = (0,1,0,-1)$. Then

$$\lambda'\Omega_n(\theta)\Big|_{\theta_3=0}\lambda = (1/n)\sum_{t=1}^{n}[\lambda'(\partial/\partial\theta)f(x_t,\theta)\Big|_{\theta_3=0}]^2 = 0.$$

Since zero for every n, $\lambda'[\lim_{n\to\infty}\Omega_n(\theta)\Big|_{\theta_3=0}]\lambda = 0$ by continuity of $\lambda'A\lambda$ in A.

Recall that $\{x_{3t}\}$ is independently and identically distributed according to the density $p_3(x_3)$. Being an age distribution, there is some (possibly unknown) maximum attained age c that is biologically possible. Then for any

continuous function $g(x)$ we must have $\int_0^c |g(x)| p_3(x)dx < \infty$ so that by Kolmogorov's Strong Law of Large Numbers (Tucker, 1967)

$$\lim_{n\to\infty} (1/n) \sum_{t=1}^{n} g(x_{3t}) = \int_0^c g(x)p_3(x)dx$$

Applying these facts to the treatment group we have

$$\lim_{n\to\infty} (2/n) \sum_{t \text{ odd}}^{n} [f(x_t,\theta) - f(x_t,\theta^o)]^2$$

$$= \int_0^c [f(x,\theta) - f(x,\theta^o)]^2 p_3(x_3)dx_3 \Big|_{(x_1,x_2)=(1,1)}$$

Applying them to the control group we have

$$\lim_{n\to\infty} (2/n) \sum_{t \text{ even}}^{n} [f(x_t,\theta) - f(x_t,\theta^o)]^2$$

$$= \int_0^c [f(x,\theta) - f(x,\theta^o)]^2 p_3(x_3)dx_3 \Big|_{(x_1,x_2)=(0,1)}$$

Then

$$\lim_{n\to\infty} (1/n) \sum_{t=1}^{n} [f(x_t,\theta) - f(x_t,\theta^o)]^2$$

$$= (1/2) \lim_{n\to\infty} (2/n)\Big\{ \sum_{t \text{ odd}}^{n} [f(x_t,\theta) - f(x_t,\theta^o)]^2 + \sum_{t \text{ even}}^{n} [f(x_t,\theta) - f(x_t,\theta^o)]^2 \Big\}$$

$$= (1/2) \sum_{(x_1,x_2)=(0,1)}^{(1,1)} \int_0^c [f(x,\theta) - f(x,\theta^o)]^2 p_3(x_3)dx_3.$$

Suppose we let $F_{12}(x_1,x_2)$ be the distribution function corresponding to the discrete density

$$p_{12}(x_1,x_2) = \begin{cases} 1/2 & (x_1,x_2) = (0,1) \\ 1/2 & (x_1,x_2) = (1,1) \end{cases}$$

and we let $F_3(x_3)$ be the distribution function corresponding to $p_3(x)$. Let $\mu(x) = F_{12}(x_1,x_2)F_3(x_3)$ then

$$\int[f(x,\theta) - f(x,\theta^o)]^2 d\mu(z) = (1/2) \sum_{(x_1,x_2)=(0,1)}^{(1,1)} \int_0^c [f(x,\theta) - f(x,\theta^o)]^2 p_3(x)dx$$

where the integral on the left is a Lebesque-Steiltjes integral (Royden, 1963, Ch. 12; or Tucker, 1967, Sec. 2.2). In this notation the limit can be given an integral representation

$$\lim_{n\to\infty} (1/n) \sum_{t=1}^{n} [f(x_t,\theta) - f(x_t,\theta^o)]^2 = \int[f(x,\theta) - f(x,\theta^o)]^2 d\mu(x).$$

These are the ideas behind Section 2 of Chapter 3. The advantage of the integral representation is that familiar results from integration theory can be used to deduce properties of limits. As an example: What is required of $f(x,\theta)$ such that

$$(\partial/\partial\theta) \lim_{n\to\infty} \sum_{t=1}^{n} f(x_t,\theta) = \lim_{n\to\infty} \sum_{t=1}^{n} (\partial/\partial\theta) f(x_t,\theta) \ ?$$

We find later that the existence of $b(x)$ with $|(\partial/\partial\theta)f(x,\theta)| \leq b(x)$ and $\int b(x)d\mu(x) < \infty$ is enough given continuity of $(\partial/\partial\theta)f(x,\theta)$.

Our last task is to verify that

$$s(\theta) = \int[f(x,\theta) - f(x,\theta^o)]^2 d\mu(x)$$

$$= (1/2) \sum_{(x_1,x_2)=(0,1)}^{(1,1)} \int_0^c [f(x,\theta) - f(x,\theta^o)]^2 p_3(x_3)dx_3$$

$$= (1/2)\int_0^c [(\theta_2 - \theta_2^o) + \theta_4 e^{\theta_3 x} - \theta_4^o e^{\theta_3^o x}]^2 p_3(x)dx$$

$$+ (1/2)\int_0^c [(\theta_1 - \theta_1^o) + (\theta_2 - \theta_2^o) + \theta_4 e^{\theta_3 x} - \theta_4^o e^{\theta_3^o x}]^2 p_3(x)dx$$

has a unique minimum. Since $s(\theta) > 0$ in general and $s(\theta^o) = 0$, the question is: Does $s(\theta) = 0$ imply that $\theta = \theta^o$? One first notes that $\theta_3^o = 0$ or $\theta_4^o = 0$ must be ruled out as in the former case any θ with $\theta_3 = 0$ and $\theta_2 + \theta_4 = \theta_2^o + \theta_4^o$ will have $s(\theta) = 0$ and in the latter case any θ with $\theta_1 = \theta_1^o$, $\theta_2 = \theta_2^o$, $\theta_4 = 0$ will have $s(\theta) = 0$. Then assume that $\theta_3^o \neq 0$ and $\theta_4^o \neq 0$ and recall that $p_3(x) > 0$ on $[0,b]$. Now $s(\theta) = 0$ implies

$$\theta_2 - \theta_2^o + \theta_4 e^{\theta_3 x} - \theta_4^o e^{\theta_3^o x} = 0 \qquad 0 \leq x \leq b$$

Differentiating we have

$$\theta_3 \theta_4 e^{\theta_3 x} - \theta_3^o \theta_4^o e^{\theta_3^o x} = 0 \qquad 0 \leq x \leq b$$

Putting $x = 0$ we have $\theta_3 \theta_4 = \theta_3^o \theta_4^o$ whence

$$e^{(\theta_3 - \theta_3^o)x} = 1 \qquad 0 \leq x \leq b$$

which implies $\theta_3 = \theta_3^o$. We now have that

$$s(\theta) = 0,\ \theta_3^o \neq 0,\ \theta_4^o \neq 0 \Rightarrow \theta_3 = \theta_3^o,\ \theta_4 = \theta_4^o.$$

But if $\theta_3 = \theta_3^o$, $\theta_4 = \theta_4^o$, and $s(\theta) = 0$ then

$$s(\theta) = (1/2)(\theta_2 - \theta_2^o)^2 + (1/2)[(\theta_1 - \theta_1^o) + (\theta_2 - \theta_2^o)]^2 = 0$$

which implies $\theta_1 = \theta_1^o$ and $\theta_2 = \theta_2^o$. In summary

$$s(\theta) = 0,\ \theta_3^o \neq 0,\ \theta_4^o \neq 0 \Rightarrow \theta = \theta^o. \quad \blacksquare$$

As seen from Example 1, checking the Identification Condition and Rank Qualification is a tedious chore to be put to at every instance one uses nonlinear methods. Uniqueness depends on the interaction of $f(x,\theta)$ and $\mu(x)$ and verification is ad hoc. Similarly for the Rank Qualification (Problem 2). As a practical matter, one should be on guard against obvious problems and can usually trust that numerical difficulties in computing $\hat{\theta}$ will serve as a sufficient warning against subtle problems as seen in the next section.

An appropriate question is how accurate are probability statements based on the asymptotic properties of nonlinear least squares estimators in applications. Specifically one might ask: How accurate are probability statements obtained by using the critical points of the t-distribution with n-p degrees of freedom to approximate the sampling distribution of

$$\tilde{t}_i = (\theta_i - \theta_i^o)/\sqrt{s^2 \hat{c}_{ii}} \quad ?$$

Monte Carlo evidence on this point is presented below using Example 1. We shall accumulate such information as we progress.

EXAMPLE 1 (continued). Table 3 shows the empirical distribution of $\tilde{t}_i$ computed from five thousand Monte Carlo trials evaluated at the critical

Table 3. Empirical Distribution of $\tilde{t}_i$ Compared to the t-distribution

Tabular Values		Empirical Distribution				
c	P(t < c)	$P(\tilde{t}_1 < c)$	$P(\tilde{t}_2 > c)$	$P(\tilde{t}_3 < c)$	$P(\tilde{t}_4 < c)$	Std. Error
-3.707	.0005	.0010	.0010	.0000	.0002	.0003
-2.779	.0050	.0048	.0052	.0018	.0050	.0010
-2.056	.0250	.0270	.0280	.0140	.0270	.0022
-1.706	.0500	.0522	.0540	.0358	.0494	.0031
-1.315	.1000	.1026	.1030	.0866	.0998	.0042
-1.058	.1500	.1552	.1420	.1408	.1584	.0050
-0.856	.200	.2096	.1900	.1896	.2092	.0057
-0.684	.2500	.2586	.2372	.2470	.2638	.0061
0.0	.5000	.5152	.4800	.4974	.5196	.0071
0.684	.7500	.7558	.7270	.7430	.7670	.0061
0.856	.8000	.8072	.7818	.7872	.8068	.0057
1.058	.8500	.8548	.8362	.8346	.8536	.0050
1.315	.9000	.9038	.8914	.8776	.9004	.0042
1.706	.9500	.9552	.9498	.9314	.9486	.0031
2.056	.9750	.9772	.9780	.9584	.9728	.0022
2.779	.9950	.9950	.9940	.9852	.9936	.0010
3.707	.9995	.9998	.9996	.9962	.9994	.0003

points of the t-distribution. The responses were generated using the inputs of Table 1 with the parameters of the model set at

$$\theta^{o} = (0,\ 1,\ -1,\ -.5)',$$
$$\sigma^2 = .001.$$

The standard errors shown in the table are the standard errors of an estimate of the probability $P(\tilde{t} < c)$ computed from 5000 Monte Carlo trials assuming that $\tilde{t}$ follows the t-distribution. If that assumption is correct, the Monte Carlo estimate of $P[\tilde{t} < c]$ follows the binomial distribution and has variance $P(t < c) \cdot P(t > c)/5000$.

Table 3 indicates that the critical points of the t-distribution describe the sampling behavior of $\tilde{t}_i$ reasonably well. For example, the Monte Carlo estimate of the Type I error for a two-tailed test of H: $\theta_3^o = -1$ using the tabular values $\pm$ 2.056 is .0556 with a standard error of .0031. Thus it seems that the actual level of the test is close enough to its nominal level of .05 for any practical purpose. However, in the next chapter we will encounter instances where this is definitely not the case.

PROBLEMS

1. Show that H: $\theta_4^o = 0$ will violate the Rank Qualification in Example 1.

2. Show that $\Omega = \lim_{n\to\infty} (1/n)F'(\theta)F(\theta)$ has full rank in Example 1 if $\theta_3^o \neq 0$ and $\theta_4^o \neq 0$.

4. METHODS OF COMPUTING LEAST SQUARES ESTIMATORS

The more widely used methods of computing nonlinear least squares estimators are Hartley's (1961) modified Gauss-Newton method and the Levenberg (1944)-Marquardt (1963) algorithm.

The Gauss-Newton method is based on the substitution of a first order Taylor's series aproximation to $f(\theta)$ about a trial parameter value θ_T in the formula for the residual sum of squares $SSE(\theta)$. The approximating sum of squares surface thus obtained is

$$SSE_T(\theta) = \|y - f(\theta_T) - F(\theta_T)(\theta - \theta_T)\|^2.$$

The value of the parameter minimizing the approximating sum of squares surface is (Problem 1)

$$\theta_M = \theta_T + [F'(\theta_T)F(\theta_T)]^{-1}F'(\theta_T)[y - f(\theta_T)].$$

It would seem that θ_M should be a better approximation to the least squares estimator $\hat{\theta}$ than θ_T in the sense that $SSE(\theta_M) < SSE(\theta_T)$. These ideas are displayed graphically in Figure 1 in the case that θ is univariate ($p=1$).

As suggested by Figure 1, $SSE_T(\theta)$ is tangent to the curve $SSE(\theta)$ at the point θ_T. The approximation is first order in the sense that one can show that (Problem 2)

$$\lim_{\|\theta-\theta_T\|\to 0} |SSE(\theta) - SSE_T(\theta)|/\|\theta - \theta_T\| = 0$$

but not second order since the best one can show in general is that

(Problem 2)

$$\lim_{\delta \to 0} \sup_{\|\theta - \theta_T\| < \delta} |SSE(\theta) - SSE_T(\theta)| / \|\theta - \theta_T\|^2 < \infty.$$

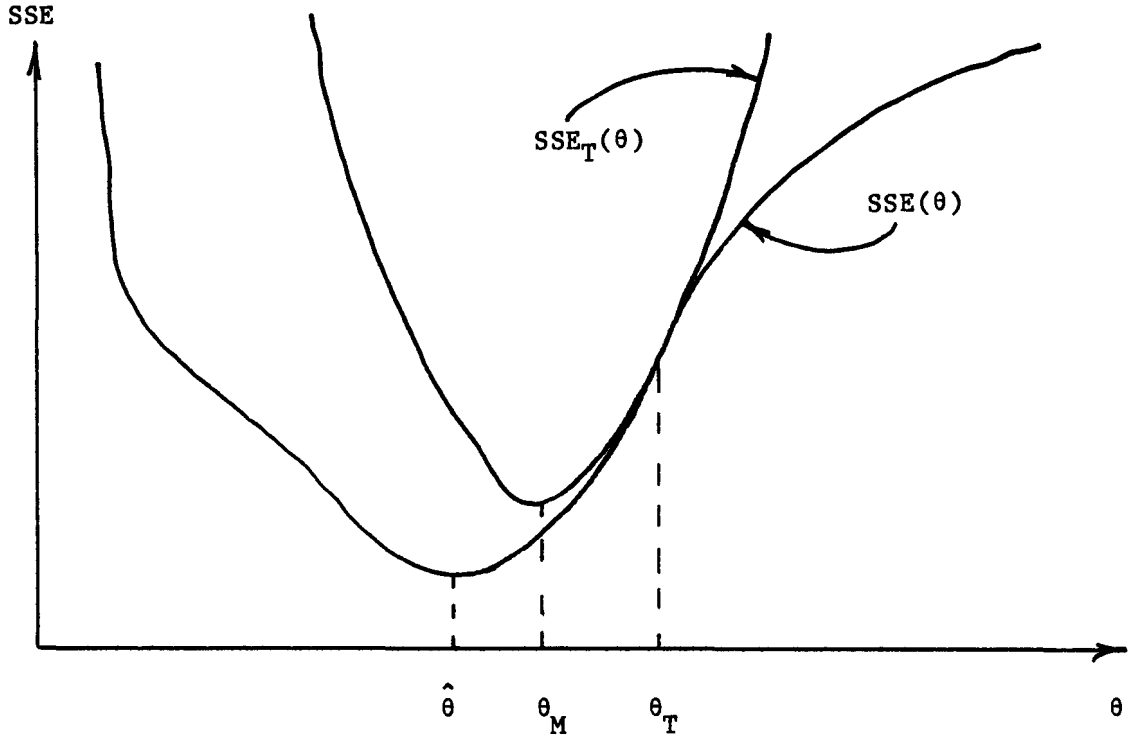


Figure 1. The Linearized Approximation to the Residual Sum of Squares Surface, an Adequate Approximation

It is not necessarily true that θ_M is closer to $\hat{\theta}$ than θ_T in the sense that $SSE(\theta_M) < SSE(\theta_T)$. This situation is depicted in Figure 2.

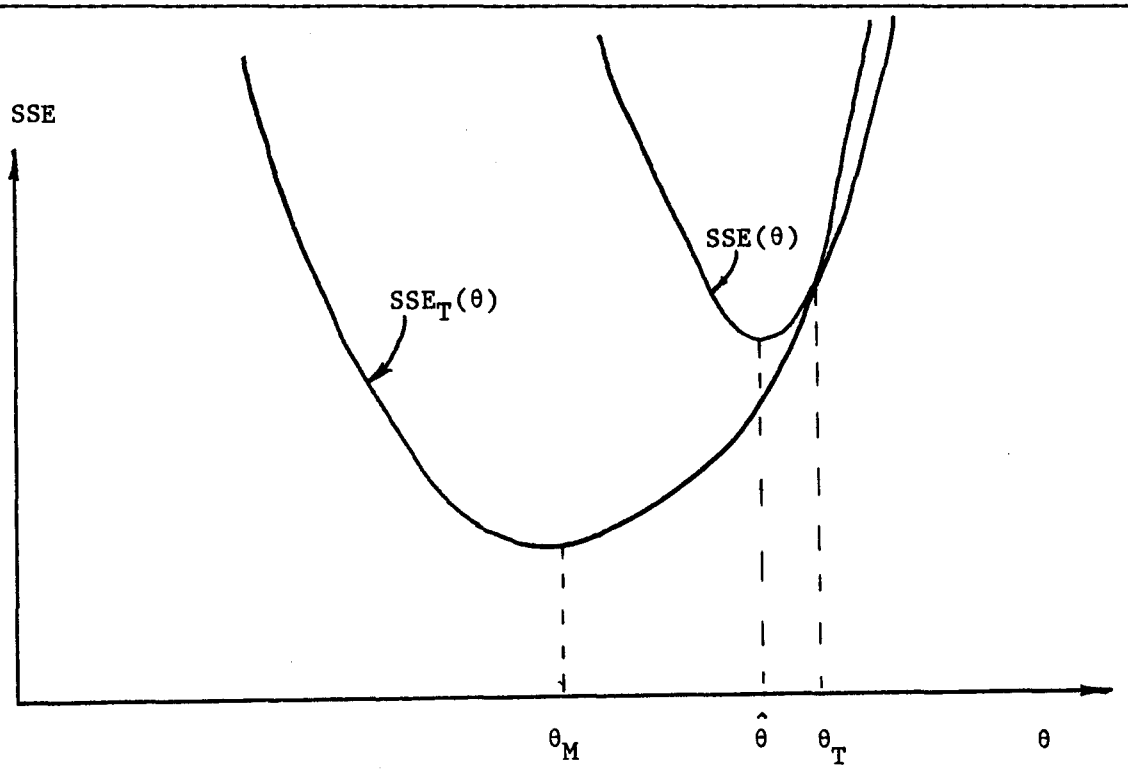


Figure 2. The Linearized Approximation to the Residual Sum of Squares Surface, A Poor Approximation

But as suggested by Figure 2, points on the line segment joining θ_T to θ_M that are sufficiently close to θ_T ought to lead to improvement. This is the case and one can show (Problem 3) that there is a λ^* such that all points with

$$\theta = \theta_T + \lambda(\theta_M - \theta_T) \qquad 0 < \lambda < \lambda^*$$

satisfy

$$SSE(\theta) < SSE(\theta_T)$$

These are the ideas that motivate the modified Gauss-Newton algorithm which is as follows:

0) Choose a starting estimate θ_o. Compute

$$D_o = [F'(\theta_o)F(\theta_o)]^{-1}F'(\theta_o)[y - f(\theta_o)].$$

Find a λ_o between 0 and 1 such that

$$SSE(\theta_o + \lambda_o D_o) < SSE(\theta_o).$$

1) Let $\theta_1 = \theta_o + \lambda_o D_o$. Compute

$$D_1 = [F'(\theta_1)F(\theta_1)]^{-1}F'(\theta_1)[y - f(\theta_1)].$$

Find a λ_1 between 0 and 1 such that

$$SSE(\theta_1 + \lambda_1 D_1) < SSE(\theta_1).$$

2) Let $\theta_2 = \theta_1 + \lambda_1 D_1$

.

.

.

There are several methods for choosing the step length λ_i at each iteration of which the simplest is to accept the first λ in the sequence

$$1, .9, .8, .7, .6, 1/2, 1/4, 1/8, \ldots$$

for which

$$SSE(\theta_i + \lambda D_i) < SSE(\theta_i)$$

as the step length λ_i. This simple approach is nearly always adequate in applications. Hartley (1961) suggests two alternative methods in his article. Gill, Murray, and Wright (1981, Sec. 4.3.2.1) discuss the problem in general from a practical point of view and follow the discussion with an annotated bibliography of recent literature. Whatever rule is used, it is essential that the computer program verify that $SSE(\theta_i + \lambda_i D_i)$ is smaller than $SSE(\theta_i)$ before taking the next iterative step. This caveat is necessry, when, for example, Hartley's quadratic interpolation formula is used to find λ_i.

The iterations are continued until terminated by a stopping rule such as

$$\|\theta_i - \theta_{i+1}\| < \varepsilon(\|\theta_i\| + \tau)$$

and

$$|SSE(\theta_i) - SSE(\theta_{i+1})| < \varepsilon\,[SSE(\theta_i) + \tau]$$

where $\varepsilon > 0$ and $\tau > 0$ are preset tolerances. Common choices are $\varepsilon = 10^{-5}$ and $\tau = 10^{-3}$. A more conservative (and costly) approach is to allow the iterations to continue until the requisite step size λ_i is so small that the fixed word length of the machine prevents differentiation between the values of $SSE(\theta_i + \lambda_i D_i)$ and $SSE(\theta_i)$. This happens sooner than one might expect and,

unfortunately, sometimes before the correct answer is obtained. Gill, Murray, and Wright (1981, Sec. 8.2.3) discuss termination criteria in general and follow the discussion with an annotated bibliography of recent literature.

Much more difficult than deciding when to stop the iterations is determining where to start them. The choice of starting values is pretty much an ad hoc process. They may be obtained from prior knowledge of the situation, inspection of the data, grid search, or trial and error. A general method of finding starting values is given by Hartley and Booker (1965). Their idea is to cluster the independent variables $\{x_t\}$ into p groups

$$x_{ij} \qquad j=1,2,\ldots,n_i; \quad i=1,2,\ldots,p$$

and fit the model

$$\bar{y}_i = \bar{f}_i(\theta) + \bar{e}_i$$

where

$$\bar{y}_i = (1/n_i) \sum_{j=1}^{n_i} y_{ij}$$

$$\bar{f}_i(\theta) = (1/n_i) \sum_{j=1}^{n_i} f(x_{ij},\theta)$$

for i=1,2,...,p. The hope is that one can find a value θ_o that solves the equations

$$\bar{y}_i = \bar{f}_i(\theta) \qquad i=1,2,\ldots,p$$

exactly. The only reason for this hope is that one has a system of p equations in p unknowns but as the system is not a linear system there is no guarantee. If an exact solution cannot be found, it is hard to see why one is better off with this new problem than with the orginal least squares problem

$$\text{minimize:} \quad SSE(\theta) = (1/n) \sum_{t=1}^{n} [y_t - f(x_t,\theta)]^2.$$

A simpler variant of their idea, and one that is much easier to use with a statistical package, is to select p representative inputs x_{t_i} with corresponding responses y_{t_i} then solve the system of nonlinear equations

$$y_{t_i} = f(x_{t_i},\theta) \qquad i=1,2,\ldots,p$$

for θ. The solution is used as the starting value. Even if iterative methods must be employed to obtain the solution it is still a viable technique since the correct answer can be recognized when found. This is not the case in an attempt to minimize $SSE(\theta)$ directly. As with Hartley-Booker, the method fails when there is no solution to the system of nonlinear equations. There is also a risk that this technique can place the starting value near a slight depression in the surface $SSE(\theta)$ and cause convergence to a local minimum that is not the global minimum. It is sound practice to try a few perturbations of θ_o as starting values and see if convergence to the same point occurs each time. We illustrate these techniques with Example 1.

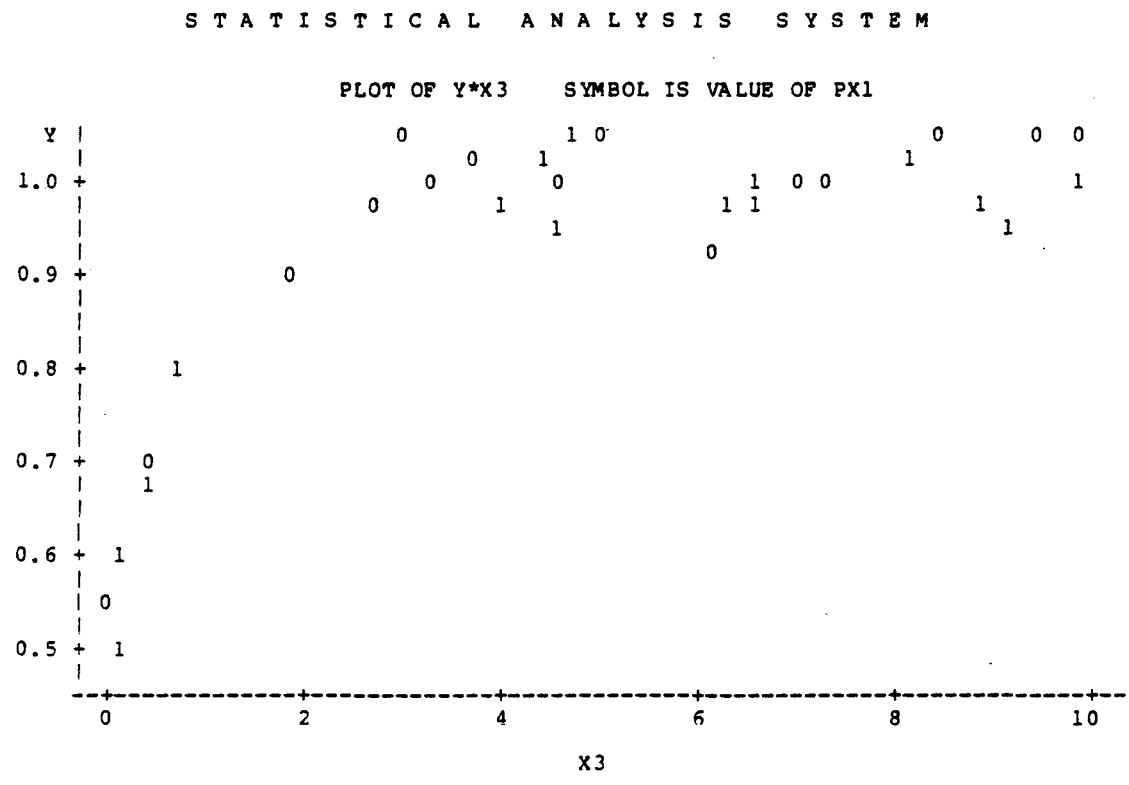
EXAMPLE 1 (continued). We begin by plotting the data as shown in Figure 3. A "1" indicates the observation is in the treatment group and a "0" indicates that the observation is in the control group. Looking at the plot, the treatment effect appears to be negligible; a starting value of zero for

Figure 3. Plot of the Data of Example 1.

SAS Statements:

```
DATA WORK01;  SET EXAMPLE1;
PX1='0';  IF X1=1 THEN PX1='1';
PROC PLOT DATA=WORK01;
PLOT Y*X3=PX1 / HAXIS = 0 TO 10 BY 2  VPOS = 24;
```

Output:



θ_1, seems reasonable. The overall impression is that the curve is concave and increasing. That is, it appears that

$$(\partial/\partial x_3)\ f(x,\theta) > 0,$$

and

$$(\partial^2/\partial x_3^2)\ f(x,\theta) < 0.$$

Since

$$(\partial/\partial x_3)\ f(x,\theta) = \theta_3\theta_4 e^{\theta_3 x_3} > 0$$

and

$$(\partial^2/\partial x_3^2)\ f(x,\theta) = \theta_3^2\theta_4 e^{\theta_3 x_3} < 0$$

we see that both θ_3 and θ_4 must be negative. Experience with exponential models suggests that what is important is to get the algebraic signs of the starting values of θ_3 and θ_4 correct and that, within reason, getting the correct magnitudes is not that important. Accordingly, take -1 as the starting value of both θ_3 and θ_4. Again, experience indicates that the starting values for parameters that enter the model linearly such as θ_1 and θ_2 are almost irrelevant, within reason, so take zero as the starting value of θ_2. In summary, inspection of a plot of the data suggests that

$$\theta = (0,\ 0,\ -1,\ -1)'$$

is a reasonable starting value.

Let us use the idea of solving equations

$$y_{t_i} = f(x_{t_i}, \theta) \qquad i=1,2,\ldots,p$$

for some representative set of inputs

$$x_{t_i} \qquad i=1,2,\ldots,p$$

to refine these visual impressions and get better starting values. We can solve the equations by minimizing

$$\sum_{i=1}^{p} [y_{t_i} - f(x_{t_i}, \theta)]^2$$

using the modified Gauss-Newton method. If the equations have a solution then the starting value we seek will produce a residual sum of squares of zero. The equation for observations in the control group ($x_1 = 0$) is

$$f(x,\theta) = \theta_2 + \theta_4 e^{\theta_3 x_3}$$

If we take two extreme values of x_3 and one where the curve is bending we should get a good fix on values for θ_2, θ_3, θ_4. Inspecting Table 1, let us select

$$x_{14} = (0,\ 1,\ 0.07)',$$
$$x_6 = (0,\ 1,\ 1.82)',$$
$$x_2 = (0,\ 1,\ 9.86)'.$$

The equation for an observation in the treatment group ($x_1 = 1$) is

Figure 4. Computation of Starting Values for Example 1.

SAS Statements:

```
DATA WORK01;  SET EXAMPLE1;
IF T=2 OR T=6 OR T=11 OR T=14 THEN OUTPUT; DELETE;
PROC NLIN DATA=WORK01 METHOD=GAUSS ITER=50 CONVERGENCE=1.0E-5;
PARMS T1=0 T2=0 T3=-1 T4=-1;
MODEL Y=T1*X1+T2*X2+T4*EXP(T3*X3);
DER.T1=X1; DER.T2=X2; DER.T3=T4*X3*EXP(T3*X3); DER.T4=EXP(T3*X3);
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

NON-LINEAR LEAST SQUARES ITERATIVE PHASE

DEPENDENT VARIABLE: Y METHOD: GAUSS-NEWTON

ITERATION	T1 T4	T2	T3	RESIDUAL SS
0	0.000000E+00 -1.00000000	0.000000E+00	-1.00000000	5.39707160
1	-0.04866000 -0.51074741	1.03859589	-0.82674151	0.00044694
2	-0.04866000 -0.51328803	1.03876874	-0.72975636	0.00000396
3	-0.04866000 -0.51361959	1.03883445	-0.73786415	0.00000000
4	-0.04866000 -0.51362269	1.03883544	-0.73791851	0.00000000
5	-0.04866000 -0.51362269	1.03883544	-0.73791852	0.00000000

NOTE: CONVERGENCE CRITERION MET.

$$f(x,\theta) = \theta_1 + \theta_2 + \theta_4 e^{\theta_3 x_3}.$$

If we can find an observation in the treatment group with an x_3 near one of the x_3's that we have already chosen then we should get a good fix on θ_1 that is independent of whatever blunders we make in guessing θ_2, θ_3, and θ_4. The eleventh observation is ideal

$$x_{11} = (1,\ 1,\ 9.86)'.$$

Figure 4 displays SAS code for selecting the subsample x_2, x_6, x_{11}, x_{14} from the original data set and solving the equations

$$y_t = f(x_t,\theta) \qquad t=2,6,11,14$$

by minimizing

$$\sum_{t=2,6,11,14} [y_t - f(x_t,\theta)]^2$$

using the modified Gauss-Newton method from a starting value of

$$\theta = (0,\ 0,\ -1,\ -1).$$

The solution is

$$\hat{\theta} = \begin{pmatrix} -0.04866 \\ 1.03884 \\ -0.73792 \\ -0.51362 \end{pmatrix}$$

Figure 5a. Example 1 Fitted by the Modified Gauss-Newton Method.

SAS Statements:

```
PROC NLIN DATA=EXAMPLE1 METHOD=GAUSS ITER=50 CONVERGENCE=1.0E-13;
PARMS T1=-0.04866 T2=1.03884 T3=-0.73792 T4=-0.51362;
MODEL Y=T1*X1+T2*X2+T4*EXP(T3*X3);
DER.T1=X1; DER.T2=X2; DER.T3=T4*X3*EXP(T3*X3); DER.T4=EXP(T3*X3);
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

NON-LINEAR LEAST SQUARES ITERATIVE PHASE

DEPENDENT VARIABLE: Y METHOD: GAUSS-NEWTON

ITERATION	T1 T4	T2	T3	RESIDUAL SS
0	-0.04866000 -0.51362000	1.03884000	-0.73792000	0.05077531
1	-0.02432899 -0.49140162	1.00985922	-1.01571093	0.03235152
2	-0.02573470 -0.50457486	1.01531500	-1.11610448	0.03049761
3	-0.02588979 -0.50490158	1.01567999	-1.11568229	0.03049554
4	-0.02588969 -0.50490291	1.01567966	-1.11569767	0.03049554
5	-0.02588970 -0.50490286	1.01567967	-1.11569712	0.03049554
6	-0.02588970 -0.50490286	1.01567967	-1.11569714	0.03049554

NOTE: CONVERGENCE CRITERION MET.

S T A T I S T I C A L A N A L Y S I S S Y S T E M 2

NON-LINEAR LEAST SQUARES SUMMARY STATISTICS DEPENDENT VARIABLE Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	4	26.34594211	6.58648553
RESIDUAL	26	0.03049554	0.00117291
UNCORRECTED TOTAL	30	26.37643764	
(CORRECTED TOTAL)	29	0.71895291	

PARAMETER	ESTIMATE	ASYMPTOTIC STD. ERROR	ASYMPTOTIC 95 % CONFIDENCE INTERVAL LOWER	UPPER
T1	-0.02588970	0.01262384	-0.05183816	0.00005877
T2	1.01567967	0.00993793	0.99525213	1.03610721
T3	-1.11569714	0.16354199	-1.45185986	-0.77953442
T4	-0.50490286	0.02565721	-0.55764159	-0.45216413

ASYMPTOTIC CORRELATION MATRIX OF THE PARAMETERS

	T1	T2	T3	T4
T1	1.000000	-0.627443	-0.085786	-0.136140
T2	-0.627443	1.000000	0.373492	-0.007261
T3	-0.085786	0.373492	1.000000	0.561533
T4	-0.136140	-0.007261	0.561533	1.000000

SAS code using this as the starting value for computing the least squares estimator with the modified Gauss-Newton method is shown in Figure 5a together with the resulting output. The least squares estimator is

$$\hat{\theta} = \begin{pmatrix} -0.02588970 \\ 1.01567967 \\ -1.115769714 \\ -0.50490286 \end{pmatrix}$$

The residual sum of squares is

$$SSE(\hat{\theta}) = 0.03049554$$

and the variance estimate is

$$s^2 = SSE(\hat{\theta})/(n-p) = 0.00117291.$$

As seen from Figure 5a, SAS prints estimated standard errors $\hat{\sigma}_i$ and correlations $\hat{\rho}_{ij}$. To recover the matrix $s^2\hat{C}$ one uses the formula:

$$s^2\hat{c}_{ij} = (\hat{\sigma}_i)(\hat{\sigma}_j)(\hat{\rho}_{ij}).$$

For example,

$$s^2c_{12} = (0.01262384)(0.00993793)(-0.627443)$$
$$= -0.000078716.$$

Figure 5b. The Matrices $s^2 C$ and C for Example 1.

$s^2 C$

	COL 1	COL 2	COL 3	COL 4
ROW 1	0.00015936	-7.8716D-05	-0.00017711	-4.4095D-05
ROW 2	-7.8716D-05	9.8762D-05	0.00060702	-1.8514D-06
ROW 3	-0.00017711	0.00060702	0.026746	0.00235621
ROW 4	-4.4095D-05	-1.8514D-06	0.00235621	0.00065829

C

	COL 1	COL 2	COL 3	COL 4
ROW 1	0.13587	-0.067112	-0.15100	-0.037594
ROW 2	-0.067112	0.084203	0.51754	-0.00157848
ROW 3	-0.15100	0.51754	22.8032	2.00887
ROW 4	-0.037594	-0.00157848	2.00887	0.56125

The matrices $s^2\hat{C}$ and $\hat{C}$ are shown in Figure 5b. ∎

The obvious approach to finding starting values is grid search. When looking for starting values by a grid search, it is only necessary to search with respect to those parameters which enter the model nonlinearly. The parameters which enter the model linearly can be estimated by ordinary multiple regression methods once the nonlinear parameters are specified. For example, once θ_3 is specified the model

$$y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_4 e^{\theta_3 x_{2t}} + e_t$$

is linear in the remaining parameters θ_1, θ_2, θ_4 and these can be estimated by linear least squares. The surface to be inspected for a minimum with respect to grid values of the parameters entering nonlinearly is the residual sum of squares after fitting for the parameters entering linearly. The trial value of the nonlinear parameters producing the minimum over the grid together with the corresponding least squares estimates of the parameters entering the model is the starting value. Some examples of plots of this sort are found toward the end of this section.

The surface to be examined for a minimum is usually locally convex. This fact can be exploited in the search to eliminate the necessity of evaluating the residual sum of squares at every point in the grid. Often, a direct search with respect to the parameters entering the model nonlinearly which exploits convexity is competitive in cost and convenience with either Hartley's or Marquardt's methods. The only reason to use the latter methods in such situations would be to obtain the matrix $[F'(\hat{\theta})F(\hat{\theta})]^{-1}$, which is printed by most implementations of either algorithm.

Of course, these same ideas can be exploited in designing an algorithm. Suppose that the model is of the form

$$f(\rho,\beta) = A(\rho)\beta$$

where ρ denotes the parameters entering nonlinearly, $A(\rho)$ is an n by K matrix, and β is a K-vector denoting the parameters entering linearly. Given ρ, the minimizing value of β is

$$\hat{\beta} = [A'(\rho)A(\rho)]^{-1}A'(\rho)y.$$

The residual sum of squares surface after fitting the parameters entering linearly is

$$SSE(\rho) = \{y - A(\rho)[A'(\rho)A(\rho)]^{-1}A'(\rho)y\}'\{y - A(\rho)[A'(\rho)A(\rho)]^{-1}A'(\rho)y\}.$$

To solve this minimization problem one can simply view

$$f(\rho) = A(\rho)[A'(\rho)A(\rho)]^{-1}A'(\rho)y$$

as a nonlinear model to be fitted to y and use, say, the modified Gauss-Newton method. Of course computing

$$(\partial/\partial\rho)\{A(\rho)[A'(\rho)A(\rho)]^{-1}A'(\rho)y\}$$

is not a trivial task but it is possible. Golub and Pereya (1973) obtain an analytic expression for $(\partial/\partial\rho)f(\rho)$ and present an algorithm exploiting it that is probably the best of its genre.

Marquardt's algorithm is similar to the Gauss-Newton method in the use of the sum of squares $SSE_T(\theta)$ to approximate $SSE(\theta)$. The difference between the two methods is that Marquardt's algorithm uses a ridge regression improvement of the approximating surface

$$\theta_\delta = \theta_T + [F'(\theta_T)F(\theta_T) + \delta I]^{-1}F'(\theta_T)[y - f(\theta_T)]$$

instead of the minimizing value θ_M. For all δ sufficiently large θ_δ is an improvement over θ_T ($SSE(\theta_\delta)$ is smaller than $SSE(\theta_T)$) under appropriate conditions (Marquardt, 1963). This fact forms the basis for Marquardt's algorithm.

The algorithm actually recommended by Marquardt differs from that suggested by this theoretical result in that a diagonal matrix S with the same diagonal elements as $F'(\theta_T)F(\theta_T)$ is substituted for the identity matrix in the expression for θ_δ. Marquardt gives the justification for this deviation in his article and, also, a set of rules for choosing δ at each iterative step. See Osborne (1972) for additional comments on these points.

Newton's method (Gill, Murray, and Wright, 1981, Sec.4.4) is based on second order Taylor's series approximation to $SSE(\theta)$ at the point θ_T;

$$SSE(\theta) \doteq SSE(\theta_T) + [(\partial/\partial\theta')SSE(\theta_T)](\theta - \theta_T)$$

$$+ \frac{1}{2}(\theta - \theta_T)'[(\partial^2/\partial\theta\partial\theta')SSE(\theta_T)](\theta - \theta_T).$$

The value of θ that minimizes this expression is

$$\theta_M = \theta_T + [-(\partial^2/\partial\theta\partial\theta')SSE(\theta_T)]^{-1}(\partial/\partial\theta)SSE(\theta_T).$$

As with the modified Gauss-Newton method one finds λ_T with

$$SSE[\theta_T + \lambda_T(\theta_M - \theta_T)] < SSE(\theta_T)$$

and takes $\theta = \theta_T + \lambda_T(\theta_M - \theta_T)$ as the next point in the iterative sequence. Now

$$-(\partial^2/\partial\theta\partial\theta')SSE(\theta_T)$$

$$= 2\,F'(\theta_T)F(\theta_T) - 2\sum_{t=1}^{n}\tilde{e}_t(\partial^2/\partial\theta\partial\theta')f(x_t,\theta_T)$$

where

$$\tilde{e}_t = y_t - f(x_t,\theta_T) \qquad t=1,2,\ldots,n.$$

From this expression one can see that the modified Gauss-Newton method can be viewed as an approximation to the Newton method if the term

$$\sum_{t=1}^{n}\tilde{e}_t(\partial^2/\partial\theta\partial\theta')f(x_t,\theta_T)$$

is negligible relative to the term $F'(\theta_t)F(\theta_T)$ for θ_T near $\hat{\theta}$; say, as a rule of thumb, when

$$\{\sum_{t=1}^{n}\sum_{i=1}^{p}\sum_{j=1}^{p}[\hat{e}_t(\partial^2/\partial\theta_i\partial\theta_j)f(x_t,\hat{\theta})]^2\}^{1/2}$$

is less than the smallest eigenvalue of $F'(\hat{\theta})F(\hat{\theta})$ where $\hat{e}_t = y_t - f(x_t,\hat{\theta})$. If this is not the case then one has what is known as the "large residual problem." In this instance it is considered sound practice to use the Newton method, or some other second order method, to compute the least squares estimator rather than the modified Gauss-Newton method. In most instances analytic computation of $(\partial^2/\partial\theta\partial\theta')f(x,\theta)$ is quite tedious and there is a considerable incentive to try and find some method to approximate

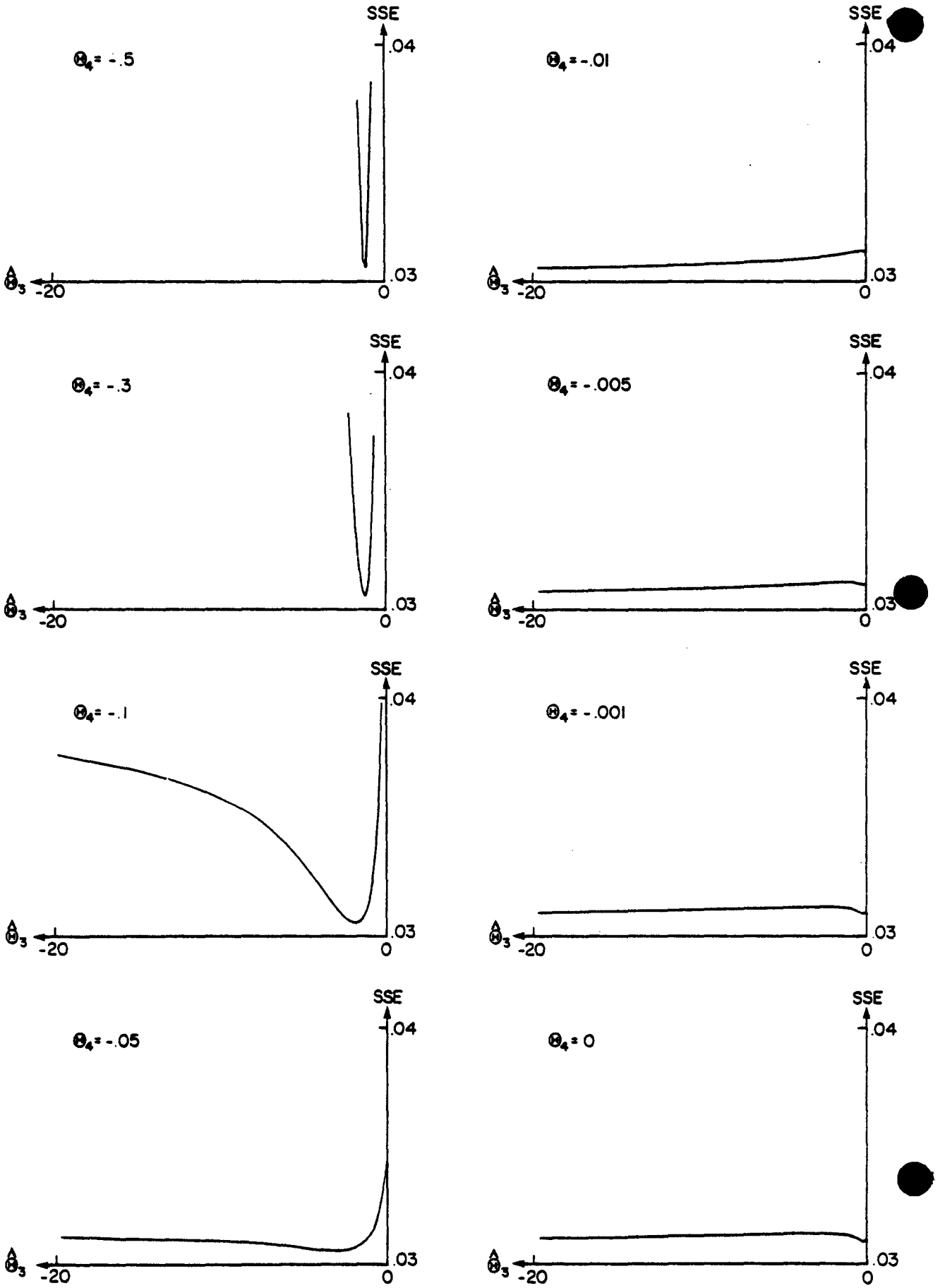
$$\sum_{t=1}^{n} \tilde{e}_t(\partial^2/\partial\theta\partial\theta')f(x_t,\theta_T)$$

without being put to this bother. The best method for doing this is probably the algorithm by Dennis, Gay and Welsch (1977).

Success, in terms of convergence to $\hat{\theta}$ from a given starting value, is not guaranteed with any of these methods. Experience indicates that failure of the iterations to converge to the correct answer depends both on the distance of the starting value from the correct answer and on the extent of over-parameterization in the response function relative to the data. These problems are interrelated in that more appropriate response functions lead to greater radii of convergence. When convergence fails, one should try to find better starting values or use a similar response function with fewer parameters. A good check on the accuracy of the numerical solution is to try several reasonable starting values and see if the iterations converge to the same answer for each starting value. It is also a good idea to plot actual responses y_t against predicted responses $\hat{y}_t = f(x_t,\hat{\theta})$; if a 45° line does not obtain then the answer is probably wrong. The following example illustrates these points.

EXAMPLE 1 (continued). Conditional on $\rho = \theta_3$, the model

Figure 6. Residual Sum of Squares Plotted Against Trial Values for θ_3 for Various True Values of θ_4.



$$f(x,\theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}$$

has three parameters $\beta = (\theta_1, \theta_2, \theta_4)$ that enter the model linearly. Then as remarked earlier, we may write

$$f(\rho) = A(\rho)[A'(\rho)A(\rho)]^{-1}A'(\rho)y$$

where a typical row of $A(\rho)$ is

$$a_t'(\rho) = (x_{1t}, x_{2t}, e^{\rho x_{3t}})$$

and treat this situation as a problem of fitting $f(\rho)$ to y by minimizing

$$SSE(\rho) = [y - f(\rho)]'[y - f(\rho)].$$

As ρ is univariate, $\hat{\rho}$ can easily be found simply by plotting $SSE(\rho)$ against ρ and inspecting the plot for the minimum. Once $\hat{\rho}$ is found,

$$\hat{\beta} = [A'(\hat{\rho})A(\hat{\rho})]^{-1}A'(\hat{\rho})y$$

gives the values of the remaining parameters.

Figure 6 shows the plots for data generated according to

$$y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_4 e^{\theta_3 x_{3t}} + e_t$$

with normally distributed errors, input variables as in Table 1, and parameter settings as in Table 3. As θ_4 is the only parameter that is varying, it

Table 4. Performance of the Modified Gauss-Newton Method

True value of θ_4[a]	Least squares estimate					Modified Gauss-Newton iterations from a start of $\hat{\theta}_i$ - .1
	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	s^2	
-.5	-.0259	1.02	-1.12	-.505	.00117	4
-.3	-.0260	1.02	-1.20	-.305	.00117	5
-.1	-.0265	1.02	-1.71	-.108	.00118	6
-.05	-.0272	1.02	-3.16	-.0641	.00117	7
-.01	-.0272	1.01	- .0452	.00758	.00120	b
-.005	-.0268	1.01	- .0971	.0106	.00119	b
-.001	-.0266	1.01	- .134	.0132	.00119	202
0	-.0266	1.01	- .142	.0139	.00119	69

[a]Parameters other than θ_4 fixed at $\theta_1 = 0$, $\theta_2 = 1$, $\theta_3 = -1$, $\sigma^2 = .001$

[b]Algorithm failed to converge after 500 iterations

serves to label the plots. The 30 errors were not regenerated for each plot, the same 30 were used each time so that θ_4 is truly all that varies in these plots.

As one sees from the various plots, fitting the model becomes an increasingly dubious proposition as $|\theta_4|$ decreases. Plots such as those in Figure 3 do not give any visual impression of an exponential trend in x_3 for $|\theta_4|$ smaller than 0.1.

Table 4 shows the deterioration in the performance of the modified Gauss-Newton method as the model becomes increasingly implausible--as $|\theta_4|$ decreases. The table was constructed by finding the local minimum nearest $\rho = 0$ ($\theta_3 = 0$) by grid search over the plots in Figure 6 and setting $\hat{\theta}_3 = \hat{\rho}$ and $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_4) = \hat{\beta}$. From the starting value

$$ {}_{(o)}\theta_i = \hat{\theta}_i - 0.1 \qquad i=1,2,3,4 $$

an attempt was made to recompute this local minimum using the modified Gauss-Newton method and the stopping rule: Stop when two successive iterations, ${}_{(i)}\theta$ and ${}_{(i+1)}\theta$, do not differ in the fifth significant digit (properly rounded) of any component. As noted, performance deteriorates for small $|\theta_4|$.

One learns from this that problems in computing the least squares estimator will usually accompany attempts to fit models with superfluous parameters. Unfortunately one can sometimes be forced into this situation when attempting to formally test the hypothesis H: $\theta_4 = 0$. We will return to this problem in the next chapter.

PROBLEMS

1. Show that

$$SSE_T(\theta) = \|y - f(\theta_T) - F(\theta_T)(\theta - \theta_T)\|^2$$

is a quadratic function of θ with minimum

$$\theta_M = \theta_T + [F'(\theta_T)F(\theta_T)]^{-1}F'(\theta_T)[y - f(\theta_T)]$$

One can see these results at sight by applying standard linear least squares theory to the linear model $z = X\beta + e$ with $z = y - f(\theta_T) + F(\theta_T)\theta_T$, $X = F(\theta_T)$, and $\beta = \theta$.

2. Set forth regularity conditions (Taylor's theorem) such that

$$SSE(\theta) = SSE(\theta_T) + [(\partial/\partial\theta)SSE(\theta_T)]'(\theta - \theta_T)$$

$$+ \frac{1}{2}(\theta - \theta_T)'[(\partial^2/\partial\theta\partial\theta')SSE(\theta_T)](\theta - \theta_T) + o(\|\theta - \theta_T\|^3)$$

Show that

$$SSE(\theta) - SSE_T(\theta) = (\theta - \theta_T)'A(\theta - \theta_T) + o(\|\theta - \theta_T\|^3)$$

where A is a symmetric matrix. Show that $|(\theta - \theta_T)'A(\theta - \theta_T)|/\|\theta - \theta_T\|^2$ is less than the largest eigenvalue of A in absolute value, $\max|\lambda_I(A)|$. Use these facts to show that

$$\lim_{\|\theta-\theta_T\|\to 0} |SSE(\theta) - SSE_T(\theta)|/\|\theta - \theta_T\| = 0$$

and

$$\lim_{\delta\to 0} \sup_{\|\theta-\theta_T\|<\delta} |SSE(\theta) - SSE_T(\theta)|/\|\theta - \theta_T\| \leq \max|\lambda_i(A)|.$$

3. Assume that θ_T is not a stationary point of $SSE(\theta)$; that is $(\partial/\partial\theta)SSE(\theta_T) \neq 0$. Set forth regularity conditions (Taylor's theorem) such that

$$SSE[\theta_T + \lambda(\theta_M - \theta_T)]$$

$$= SSE(\theta_T) + \lambda[(\partial/\partial\theta)SSE(\theta_T)]'(\theta_M - \theta_T) + o(\lambda^2)$$

Let $F_T = F(\theta_T)$, $\hat{e}_T = [y - f(\theta_T)]$ and show that this equation reduces to

$$SSE[\theta_T + \lambda(\theta_M - \theta_T)] = SSE(\theta_T) + [-2\hat{e}_T'F_T(F_T'F_T)^{-1}F_T'\hat{e}_T + o(\lambda^2)/\lambda]\lambda$$

There must be a λ^* such that

$$[-2\hat{e}_T'F_T(F_T'F_T)^{-1}F_T'\hat{e}_T + o(\lambda^2/\lambda)] < 0$$

for all λ with $0 < \lambda < \lambda^*$, why? Thus

$$SSE[\theta_T + \lambda(\theta_M - \theta_T)] < SSE(\theta_T)$$

for all λ with $0 < \lambda < \lambda^*$.

4. (Convergence of the Modified Gauss-Newton Method). Supply the missing details in the proof of the following result.

Theorem: Let

$$Q(\theta) = \sum_{t=1}^{n} [y_t - f(x_t,\theta)]^2.$$

Conditions: There is a convex, bounded subset S of R^p and θ_o interior to S such that:

1) $(\partial/\partial\theta)f(x_t,\theta)$ exists and is continuous over $\bar{S}$ for $t = 1,2,\ldots,n$;
2) $\theta \in S$ implies the rank of $F(\theta)$ is p;
3) $Q(\theta_o) < \tilde{Q} = \inf\{Q(\theta)$: θ a boundary point of $S\}$;
4) There does not exist θ', θ'' in S such that

$$(\partial/\partial\theta)Q(\theta') = (\partial/\partial\theta)Q(\theta'') = 0 \text{ and } Q(\theta') = Q(\theta'').$$

Construction: Construct a sequence $\{\theta_\alpha\}_{\alpha=1}^{\infty}$ as follows:

0) Compute $D_0 = [F'(\theta_0)F(\theta_0)]^{-1}F'(\theta_0)[y - f(\theta_0)]$.
Find λ_0 which minimizes $Q(\theta_0 + \lambda D_0)$ over
$\Lambda_0 = \{\lambda: 0 \leq \lambda \leq 1, \theta_0 + \lambda D_0 \in \bar{S}\}$.

1) Set $\theta_1 = \theta_0 + \lambda_0 D_0$.
Compute $D_1 = [F'(\theta_1)F(\theta_1)]^{-1}F'(\theta_1)[y - f(\theta_1)]$.
Find λ_1 which minimizes $Q(\theta_1 + \lambda D_1)$ over
$\Lambda_1 = \{\lambda: 0 \leq \lambda \leq 1, \theta_1 + \lambda D_1 \in \bar{S}\}$.

2) Set $\theta_2 = \theta_1 + \lambda_1 D_1$.

.

.

.

Conclusions. Then for the sequence $\{\theta_\alpha\}_{\alpha=1}^{\infty}$ it follows that:

1) θ_α is an interior point of S for $\alpha = 1, 2, \ldots$.

2) The sequence $\{\theta_\alpha\}$ converges to a limit of θ^* which is interior to S.

3) $(\partial/\partial\theta)Q(\theta^*) = 0$.

Proof. We establish Conclusion 1. The conclusion will follow by induction if we show that θ_α interior to S and $Q(\theta_\alpha) < \tilde{Q}$ imply λ_α minimizing $Q(\theta_\alpha + \lambda D_\alpha)$ over Λ_α exists and $\theta_{\alpha+1}$ is an interior point of S. Let $\theta_\alpha \in S^o$ and consider the set

$$\hat{S} = \{\theta \in \bar{S}: \theta = \theta_\alpha + \lambda D_\alpha, 0 \leq \lambda \leq 1\}.$$

$\hat{S}$ is a closed, bounded line segment contained in $\bar{S}$, why? There is a θ' in $\hat{S}$ minimizing Q over $\hat{S}$, why? Hence, there is a λ_α $(\theta' = \theta_\alpha + \lambda_\alpha D_\alpha)$ minimizing $Q(\theta_\alpha + \lambda D_\alpha)$ over Λ_α. Now θ' is either an interior point of $\bar{S}$ or a boundary point of $\bar{S}$. By Lemma 2.2.1 of Blackwell and Girshick (1954, p. 32) S and $\bar{S}$ have the same interior points and boundary points. If θ' were a boundary point of S we would have

$$\tilde{Q} \leq Q(\theta') \leq Q(\theta_\alpha) < \tilde{Q}$$

which is not possible. Then θ' is an interior point of S. Since $\theta_{\alpha+1} = \theta'$ we have established Conclusion 1.

We establish Conclusions 2, 3. By construction $0 \leq Q(\theta_{\alpha+1}) \leq Q(\theta_\alpha)$ hence $Q(\theta_\alpha) \to Q^*$ as $\alpha \to \infty$. The sequence $\{\theta_\alpha\}$ must have a convergent subsequence $\{\theta_\beta\}_{\beta=1}^{\infty}$ with limit $\theta^* \in \bar{S}$, why? $Q(\theta_\beta) \to Q(\theta^*)$ so $Q(\theta^*) = Q^*$, why? θ^* is either an interior point of $\bar{S}$ or a boundary point. The same holds for S as we saw above. If θ^* were a boundary point of S then $\tilde{Q} \leq Q(\theta^*) \leq Q(\theta^0)$ which is impossible because $Q(\theta_0) < \tilde{Q}$. So θ^* is an interior point of S.

The function

$$D(\theta) = [F'(\theta)F(\theta)]^{-1}F'(\theta)[y - f(\theta)]$$

is continuous over S, why? Thus

$$\lim_{\beta\to\infty} D_\beta = \lim_{\beta\to\infty} D(\theta_\beta) = D(\theta^*) = D^*.$$

Suppose $D^* \neq 0$ and consider the function $q(\lambda) = Q(\theta^* + \lambda D^*)$ for $\lambda \in [-\eta, \eta]$ where $0 < \eta \leq 1$ and $\theta^* \pm \eta D^*$ are interior points of S.

$$q'(0) = (\partial/\partial\theta')Q(\theta^* + \lambda D^*)D^*|_{\lambda=0}$$

$$= (-2)[y - f(\theta^*)]'F(\theta^*)D^*$$

$$= (-2)D^{*\prime}F'(\theta^*)F(\theta^*)D^*$$

$$< 0,$$

why? Choose $\varepsilon > 0$ so that $\varepsilon < -q'(0)$. By the definition of derivative there is a $\lambda^* \in (0, \frac{1}{2}\eta)$ such that

$$Q(\theta^* + \lambda^* D^*) - Q(\theta^*) = q(\lambda^*) - q(0)$$

$$< [q'(0) + \varepsilon]\lambda^*.$$

Since Q is continuous for $\theta \in S$ we may choose $\gamma > 0$ such that $-\gamma > [q'(0) + \varepsilon]\lambda^*$ and there is $\delta > 0$ such that

$$\|\theta_\beta + \lambda^* D_\beta - \theta^* - \lambda^* D^*\| < \delta$$

implies

$$Q(\theta_\beta + \lambda^* D) - Q(\theta^* + \lambda^* D^*) < \gamma$$

Then for all β sufficiently large we have

$$Q(\theta_\beta + \lambda^* D_\beta) - Q(\theta^*) < [q'(0) + \varepsilon]\lambda^* + \gamma = -c^2.$$

Now for β large enough $\theta_\beta + \lambda^* D_\beta$ is interior to S so that $\lambda^* \in \Lambda_\beta$ and we obtain

$$Q(\theta_{\beta+1}) - Q(\theta^*) < -c^2.$$

This contradicts the fact that $Q(\theta_\beta) \to Q(\theta^*) = Q^*$ as $\beta \to \infty$; thus D^* must be the zero vector. Then it follows that

$$(\partial/\partial\theta)Q(\theta^*) = (-2)F'(\theta^*)[y - f(\theta^*)]$$

$$= (-2)F'(\theta^*)F(\theta^*)D^*$$

$$= 0.$$

Given any subsequence of $\{\theta_\alpha\}$ we have by the above that there is a convergent subsequence with limit point $\theta' \in S$ such that

$$(\partial/\partial\theta)Q(\theta') = 0 = (\partial/\partial\theta)Q(\theta^*)$$

and

$$Q(\theta') = Q^* = Q(\theta^*).$$

By Hypothesis 4, $\theta' = \theta^*$ so that $\theta_\alpha \to \theta'$ as $\alpha \to \infty$.

5. HYPOTHESIS TESTING

Assuming that the data follow the model

$$y = f(\theta^o) + e, \qquad e \sim N(0, \sigma^2 I)$$

consider testing the hypothesis

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

where $h(\theta)$ is a once continuously differentiable function mapping R^p into R^q with Jacobian

$$H(\theta) = (\partial/\partial\theta')h(\theta)$$

of order q by p. When $H(\theta)$ is evaluated at $\theta = \hat{\theta}$ we shall write $\hat{H}$,

$$\hat{H} = H(\hat{\theta}).$$

and at $\theta = \theta^o$ write H,

$$H = H(\theta^o),$$

In Chapter 4 we shall show that $h(\hat{\theta})$ may be characterized as

$$h(\hat{\theta}) = h(\theta^o) + H(F'F)^{-1}F'e + o_p(1/\sqrt{n})$$

where, recall, $F = (\partial/\partial\theta')f(\theta^o)$. Ignoring the remainder term, we have

$$h(\hat{\theta}) \dot{\sim} N_q[h(\theta^o),\ \sigma^2 H(F'F)^{-1}H']$$

whence

$$h'(\hat{\theta})\ [H(F'F)^{-1}H']^{-1}h(\hat{\theta})/\sigma^2$$

is (approximately) distributed as the non-central chi-square distribution (Appendix 1) with q degrees of freedom and non-centrality parameter

$$\lambda = h'(\theta^o)[H(F'F)^{-1}H']^{-1}h(\theta^o)/(2\sigma^2).$$

Recalling that to within the order of approximation $o_p(1/n)$, $(n-p)s^2/\sigma^2$ is distributed independently of $\hat{\theta}$ as the chi-square distribution with n-p degrees of freedom we have (approximately) that the ratio

$$\frac{h'(\hat{\theta})[H(F'F)^{-1}H']^{-1}h(\hat{\theta})/(q\sigma^2)}{(n-p)s^2/[(n-p)\sigma^2]}$$

follows the non-central F distribution (Appendix 1) with q numerator degrees of freedom, n-p denominator degrees of freedom, and non-centrality parameter λ; denoted as $F'(q,\ n-p,\ \lambda)$. Cancelling like terms in the numerator and denominator, we have

$$h'(\hat{\theta})[H(F'F)^{-1}H']^{-1}h(\hat{\theta})/(qs^2) \sim F'(q,\ n-p,\ \lambda)$$

In applications, estimates $\hat{H}$ and $\hat{C}$ must be substituted for H and $(F'F)^{-1}$

where, recall, $\hat{C} = [F'(\hat{\theta})F(\hat{\theta})]^{-1}$. The resulting statistic

$$W = h'(\hat{\theta})(\hat{H}\,\hat{C}^{-1}\,\hat{H}')^{-1}h(\hat{\theta})/(qs^2)$$

is usually called the Wald test statistic.

To summarize this discussion, the Wald test rejects the hypothesis

$$H: h(\theta^o) = 0$$

when the statistic

$$W = h'(\hat{\theta})(\hat{H}'\hat{C}\hat{H})^{-1}h(\hat{\theta})/(qs^2)$$

exceeds the upper $\alpha \times 100\%$ critical point of the F distribution with q numerator degrees of freedom and n-p denominator degrees of freedom; denoted as $F^{-1}(1-\alpha;\ q,\ n-p)$. We illustrate by example.

EXAMPLE 1 (continued). Recalling that

$$f(x,\theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}$$

consider testing the hypothesis of no treatment effect

$$H: \theta_1 = 0 \quad \text{against} \quad A: \theta_1 \neq 0.$$

For this case

$$h(\theta) = \theta_1$$

$$H(\theta) = (\partial/\partial\theta')h(\theta) = (1,0,0,0)$$

$$h(\hat{\theta}) = -0.02588970 \qquad \text{(from Figure 5a)}$$

$$\hat{H} = (\partial/\partial\theta')h(\hat{\theta}) = (1,0,0,0)$$

$$\hat{H}\hat{C}\hat{H}' = \hat{c}_{11} = 0.13587 \qquad \text{(from Figure 5b)}$$

$$s^2 = 0.00117291 \qquad \text{(from Figure 5a)}$$

$$q = 1$$

$$W = h'(\hat{\theta})(\hat{H}\hat{C}\hat{H}')^{-1}h(\theta)/(qs^2)$$

$$= (-0.02588970)(0.13587)^{-1}(-0.02588970)/(1 \times 0.00117291)$$

$$= 4.2060$$

The upper 5% critical point of the F distribution with 1 numerator degree of freedom and 26 = 30 - 4 denominator degrees of freedom is

$$F^{-1}(.95;\ 1,\ 26) = 4.22$$

so one fails to reject the null hypothesis.

Of course, in this simple instance one can compute a t-statistic directly from the output shown in Figure 5a as

$$t = (-0.02588970)/(0.01262384)$$

$$= -2.0509$$

and compare the absolute value with

$$t^{-1}(.975;\ 26) = 2.0555. \quad \blacksquare$$

In simple examples such as the proceeding, one can work directly from printed output such as Figure 5a. But anything more complicated requires some programming effort to compute and invert $\hat{H}\hat{C}\hat{H}'$. There are a variety of ways to do this; we shall describe a method that is useful pedagogically as it builds on the ideas of the previous section and is easy to use with a statistical package. It also has the advantage of saving the bother of looking up the critical values of the F distribution.

Suppose that one fits the model

$$\hat{e} = \hat{F}\beta + u$$

by least squares and tests the hypothesis

$$H: \hat{H}\beta = h(\hat{\theta}) \quad \text{against A: } \hat{H}\beta \neq h(\hat{\theta})$$

The computed F statistic will be

$$F = \frac{[\hat{H}\hat{\beta} - h(\hat{\theta})]' [\hat{H}(\hat{F}'\hat{F})^{-1}\hat{H}']^{-1} [\hat{H}\hat{\beta} - h(\hat{\theta})]/q}{[\hat{e} - \hat{F}\beta]' [\hat{e} - \hat{F}\beta]/(n-p)}$$

but since

$$0 = (\partial/\partial\theta)SSE(\hat{\theta}) = -2\hat{F}'\hat{e}$$

we have

$$0 = (\hat{F}'\hat{F})^{-1}\hat{F}'\hat{e} = \hat{\beta}$$

and the computed F statistic reduces to

$$W = h'(\hat{\theta})(\hat{H}\hat{C}\hat{H}')^{-1}h(\hat{\theta})/(qs^2).$$

Thus, any statistical package that can compute a linear regression and test a linear hypothesis becomes a convenient tool for computing the Wald test statistic. We illustrate these ideas in the next example.

EXAMPLE 1 (continued). Recalling that the response function is

$$f(x,\theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}$$

consider testing

$$H: (\partial/\partial x_3)f(x,\theta)\Big|_{x_3=1} = 1/5 \quad \text{against} \quad A: (\partial/\partial x_3)f(x,\theta)\Big|_{x_3=1} \neq 1/5$$

or equivalently

$$H: \theta_3\theta_4 e^{\theta_3} = 1/5 \quad \text{against} \quad A: \theta_3\theta_4 e^{\theta_3} \neq 1/5.$$

We have

$$h(\theta) = \theta_3\theta_4 e^{\theta_3} - 1/5$$

$$H(\theta) = (\partial/\partial\theta')h(\theta) = [0,\ 0,\ \theta_4(1 + \theta_3)e^{\theta_3},\ \theta_3 e^{\theta_3}]$$

$$h(\theta) = (-1.11569714)(-0.50490286)e^{-1.11569714} - 0.2 \qquad \text{(from Figure 5a)}$$

$$= -0.0154079303$$

$$\hat{H} = (\partial/\partial\theta')h(\hat{\theta})$$

$$= (0,\ 0,\ 0.0191420895,\ -0.365599176) \qquad \text{(from Figure 5a)}$$

$$h'(\hat{\theta})(\hat{H}\hat{C}\hat{H}')^{-1}h(\hat{\theta})/1 = 0.0042964 \qquad \text{(from Figure 7)}$$

$$s^2 = 0.001172905 \qquad \text{(from Figure 5a or 7)}$$

$$W = 3.6631 \qquad \text{(from Figure 7 or by division)}$$

Since $F^{-1}(.95;\ 1,\ 26) = 4.22$ one fails to reject at the 5% level. The p-value is 0.0667 as shown in Figure 7; that is $1 - F(3.661;\ 1,\ 26) = 0.0667$.

Also shown in Figure 7 are the computations for the previous example as well as computations for the joint hypothesis.

$$H:\ \theta_1 = 0 \ \text{ and } \ \theta_3\theta_4 e^{\theta_3} = 1/5 \ \text{ against } \ A:\ \theta_1 \neq 0 \ \text{ or } \ \theta_3\theta_4 e^{\theta_3} \neq 1/5.$$

The joint hypothesis is included to illustrate the computations for the case $q > 1$. One rejects the joint hypothesis at the 5% level; the p-value is 0.0210. ∎

We have noted in the somewhat heuristic derivation of the Wald test that W is distributed as the non-central F distribution. What can be shown rigorously (Chapter 4) is that

$$W = Y + o_p(1/n)$$

Figure 7. Illustration of Wald Test Computations with Example 1.

SAS Statements:

```
DATA WORK01;  SET EXAMPLE1;
T1=-0.02588970;  T2=1.01567967;  T3=-1.11569714;  T4=-0.50490286;
E=Y-(T1*X1+T2*X2+T4*EXP(T3*X3));
DER_T1=X1; DER_T2=X2; DER_T3=T4*X3*EXP(T3*X3); DER_T4=EXP(T3*X3);
PROC REG DATA=WORK01;  MODEL E = DER_T1 DER_T2 DER_T3 DER_T4 / NOINT;
FIRST: TEST DER_T1=0.02588970;
SECOND:TEST 0.0191420895*DER_T3-0.365599176*DER_T4=-0.0154079303;
JOINT: TEST DER_T1=0.02588970,
            0.0191420895*DER_T3-0.365599176*DER_T4=-0.0154079303;
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

DEP VARIABLE: E

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	4	3.29597E-17	8.23994E-18	0.000	1.0000
ERROR	26	0.030496	0.001172905		
U TOTAL	30	0.030496			

ROOT MSE	0.034248	R-SQUARE	0.0000
DEP MEAN	4.13616E-11	ADJ R-SQ	-0.1154
C.V.	82800642118		

NOTE: NO INTERCEPT TERM IS USED. R-SQUARE IS REDEFINED.

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
DER_T1	1	1.91639E-09	0.012624	0.000	1.0000
DER_T2	1	-6.79165E-10	0.009937927	-0.000	1.0000
DER_T3	1	1.52491E-10	0.163542	0.000	1.0000
DER_T4	1	-1.50709E-09	0.025657	-0.000	1.0000

TEST: FIRST	NUMERATOR:	.0049333	DF: 1	F VALUE:	4.2060
	DENOMINATOR:	.0011729	DF: 26	PROB >F :	0.0505
TEST: SECOND	NUMERATOR:	.0042964	DF: 1	F VALUE:	3.6631
	DENOMINATOR:	.0011729	DF: 26	PROB >F :	0.0667
TEST: JOINT	NUMERATOR:	.0052743	DF: 2	F VALUE:	4.4968
	DENOMINATOR:	.0011729	DF: 26	PROB >F :	0.0210

$$Y \sim F'(q,\ n-p,\ \lambda)$$

$$\lambda = h'(\theta^o)\{H(\theta^o)[F'(\theta^o)F(\theta^o)]^{-1}H'(\theta^o)\}^{-1}h(\theta^o)/(2\sigma^2)$$

That is, Y is distributed as the non-central F distribution with q numerator degrees of freedom, n-p denominator degrees of freedom, and non-centrality parameter λ (Appendix 1). The computation of power requires computation of λ and use of charts (Pearson and Hartley, 1951; Fox, 1956) of the non-central F distribution. One convenient source for the charts is Scheffe (1959). The computation of λ is very little different from the computation of W itself and one can use exactly the same strategy used in the previous example to obtain

$$h'(\theta^o)\{H(\theta^o)[F'(\theta^o)F(\theta^o)]^{-1}H'(\theta^o)\}^{-1}h(\theta^o)/q$$

and then multiply by $q/(2\sigma^2)$ to obtain λ. Alternatively one can write code in some programming language to compute λ. To add variety to the discussion, we shall illustrate the latter approach using PROC MATRIX in SAS.

EXAMPLE 1 (continued). Recalling that

$$f(x,\theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}$$

let us approximate the probability that the Wald test rejects the following three hypotheses at the 5% level when the true values of the parameters are

$$\theta^o = (.03,\ 1,\ -1.4,\ -.5)'$$
$$\sigma^2 = .001.$$

Figure 8. Illustration of Wald Test Power Computations with Example 1.

SAS Statements:

```
PROC MATRIX;  FETCH X DATA=EXAMPLE1(KEEP=X1 X2 X3);
T1=.03;  T2=1;  T3=-1.4;  T4=-.5;  S=.001;  N=30;
F1=X(,1);  F2=X(,2);  F3=T4*(X(,3)#EXP(T3*X(,3)));  F4=EXP(T3*X(,3));
F=F1||F2||F3||F4;  C=INV(F'*F);
SMALL_H1=T1;  H1=1 0 0 0;
LAMBDA=SMALL_H1'*INV(H1*C*H1')*SMALL_H1#/(2*S); PRINT LAMBDA;
SMALL_H2=(T3#T4#EXP(T3)-1#/5);  H2=0||0||T4#(1+T3)#EXP(T3)||T3#EXP(T3);
LAMBDA=SMALL_H2'*INV(H2*C*H2')*SMALL_H2#/(2*S); PRINT LAMBDA;
SMALL_H3=SMALL_H1//SMALL_H2;  H3=H1//H2;
LAMBDA=SMALL_H3'*INV(H3*C*H3')*SMALL_H3#/(2*S); PRINT LAMBDA;
```

Output:

```
                S T A T I S T I C A L   A N A L Y S I S   S Y S T E M          1


                LAMBDA          COL1

                ROW1          3.3343


                LAMBDA          COL1

                ROW1         5.65508


                LAMBDA          COL1

                ROW1         9.88196
```

The three null hypotheses are:

$$H_1: \theta_1 = 0,$$

$$H_2: \theta_3\theta_4 e^{\theta_3} = 1/5,$$

$$H_3: \theta_3 = 0 \quad \text{and} \quad \theta_3\theta_4 e^{\theta_3} = 1/5.$$

PROC MATRIX code to compute

$$\lambda = h'(\theta^o)\{H(\theta^o)[F'(\theta^o)F(\theta^o)]^{-1}H'(\theta^o)\}^{-1}h(\theta^o)/(2\sigma^2)$$

for each of the three cases is shown in Figure 8. We obtain

$\lambda_1 = 3.3343$ (from Figure 8)

$\lambda_2 = 5.65508$ (from Figure 8)

$\lambda_3 = 9.88196$ (from Figure 8)

Then from the Pearson-Hartley charts of the non-central F distribution in Scheffé (1959) we obtain

$$1 - F'(4.22;\ 1,\ 26,\ 3.3343) = .70,$$

$$1 - F'(4.22;\ 1,\ 26,\ 5.65508) = .90,$$

$$1 - F'(3.37;\ 2,\ 26,\ 9.88196) = .97.$$

For the first hypothesis one approximates $P(W > F_\alpha)$ by $P(Y > F_\alpha) = .70$ where

Table 5: Monte Carlo Power Estimates for the Wald Test

Parameters*		$H_o: \theta_1 = 0$ against $H_1: \theta_1 \neq 0$				$H_o: \theta_3 = -1$ against $H_1: \theta_3 \neq -1$			
				Monte Carlo				Monte Carlo	
θ_1	θ_3	λ	$P[Y > F_\alpha]$	$P[W > F_\alpha]$	STD. ERR.	λ	$P[X > F_\alpha]$	$P[W > F_\alpha]$	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

* $\theta_2 = 1$, $\theta_4 = -.5$, $\sigma^2 = .001$

$F_\alpha = F^{-1}(.95;\ 1,\ 26) = 4.22$, and so on for the other two cases.

The natural question is: How accurate are these approximations? In this instance the Monte Carlo simulations reported in Table 5 indicates that the approximation is accurate enough for practical purposes but later on we shall see examples showing fairly poor approximations to $P(W > F_\alpha)$ by $P(Y > F_\alpha)$. Table 5 was constructed by generating five thousand responses using the response function

$$f(x,\theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_4^{\theta_3 x_3}$$

and the inputs shown in Table 1. The parameters used were $\theta_2 = 1$, $\theta_4 = -.5$, and $\sigma^2 = .001$ excepting θ_1 and θ_2 which were varied as shown in Table 5. Power for a test of H: $\theta_1 = 0$ and H: $\theta_3 = -1$ is computed for $P(Y > F_\alpha)$ and compared to $P(W > F_\alpha)$ estimated from the Monte Carlo trials. The standard errors in the table refer to the fact that the Monte Carlo estimate of $P(W < F_\alpha)$ is binomially distributed with $n = 5000$ and $p = P(Y > F_\alpha)$. Thus, $P(W > F_\alpha)$ is estimated with a standard error of $\{P(Y > F_\alpha)[1 - P(Y > F_\alpha)]/5000\}^{1/2}$. These simulations are described in somewhat more detail in Gallant (1975b).

One of the most familiar methods of testing a linear hypothesis

$$H: R\beta = r \quad \text{against} \quad A: R\beta \neq r$$

for the linear model

$$y = X\beta + e$$

is: First, fit the full model by least squares obtaining

$$SSE_{full} = (y - X\hat{\beta})'(y - X\hat{\beta})$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

Second, refit the model subject to the null hypothesis that $R\beta = r$ obtaining

$$SSE_{reduced} = (y - X\tilde{\beta})'(y - X\tilde{\beta})$$

$$\tilde{\beta} = \hat{\beta} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat{\beta});$$

Third, compute the F statistic

$$F = \frac{(SSE_{reduced} - SSE_{full})/q}{(SSE_{full})/(n - p)}$$

where q is the number of restrictions on β (number of rows in R), p is the number of columns in X, and n the number of observations--full rank matrices being assumed throughout. One rejects for large values of F. If one assumes normal errors in the nonlinear model

$$y = f(\theta) + e \qquad e \sim N_p(0, \sigma^2 I)$$

and derives the likelihood ratio test statistic for the hypothesis

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

one obtains exactly the same test as just described (Problem 1). The statistic is computed as follows.

First, compute

$$\hat{\theta} \text{ minimizing } SSE(\theta) = [y - f(\theta)]'[y - f(\theta)]$$

using the methods of the previous section and let

$$SSE_{full} = SSE(\hat{\theta}).$$

Second, refit under the null hypothesis by computing

$$\tilde{\theta} \text{ minimizing } SSE(\theta) \text{ subject to } h(\theta) = 0$$

using methods discussed immediately below, and let

$$SSE_{reduced} = SSE(\tilde{\theta}).$$

Third, compute the statistic

$$L = \frac{(SSE_{reduced} - SSE_{full})/q}{(SSE_{full})/(n - p)}$$

Recall that $h(\theta)$ maps R^p into R^q so that q is, in a sense, the number of restrictions on θ. One rejects H: $h(\theta^o) = 0$ when L exceed the $\alpha \times 100\%$ critical point F_α of the F distribution with q numerator degrees of freedom and n-p denominator degrees of freedom; $F_\alpha = F^{-1}(1 - \alpha; q, n - p)$. Later on, we shall verify that L is distributed according to the F distribution if

$h(\theta^o) = 0$. For now, let us consider computational aspects.

General methods for minimizing $SSE(\theta)$ subject to $h(\theta) = 0$ are given in Gill, Murray, and Wright (1981). But it is almost always the case in practice that a hypothesis written as a parametric restriction

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

can easily be rewritten as a functional dependency

$$H: \theta^o = g(\rho) \text{ for some } \rho^o \quad \text{against} \quad A: \theta^o \neq g(\rho) \text{ for any } \rho.$$

Here ρ is an r-vector with $r = p-q$. In general one obtains $g(\rho)$ by augmenting the equations

$$h(\theta) = \tau$$

by the equations

$$\phi(\theta) = \rho$$

which are chosen such that the system of equations

$$h(\theta) = \tau$$

$$\phi(\theta) = \rho$$

is a one-to-one transformation with inverse

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

Then imposing the condition

$$\theta = \psi(\rho,0)$$

is equivalent (Problem 2) to imposing the condition

$$h(\theta) = 0$$

so that the desired functional dependency is obtained by putting

$$\theta = g(\rho).$$

But usually $g(\rho)$ can be constructed at sight on an ad hoc basis without resorting to these formalities as seen in the later examples.

The null hypothesis is that the data follow the model

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

and that θ^o satisfies

$$h(\theta^o) = 0.$$

Equivalently, the null hypothesis is that the data follow the model

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

and

$$\theta^o = g(\rho) \text{ for some } \rho^o.$$

But the latter statement can be expressed more simply as: The null hypothesis is that the data follow the model

$$y_t = f[x_t, g(\rho^o)] + e_t.$$

In vector notation,

$$y = f[g(\rho)] + e.$$

This is, of course, merely a nonlinear model that can be fitted by the methods described previously. One computes

$$\hat{\rho} \text{ minimizing } SSE[g(\rho)] = \{y - f[g(\rho)]\}'\{y - f[g(\rho)]\}$$

by, say, the modified Gauss-Newton method. Then

$$SSE_{reduced} = SSE[g(\hat{\rho})]$$

because $\tilde{\theta} = g(\hat{\rho})$ (Problem 3).

The fact that $f[x,g(\rho)]$ is a composite function gives derivatives some structure that can be exploited in computations. Let

$$G(\rho) = (\partial/\partial\rho')g(\rho),$$

that is, $G(\rho)$ is the Jacobian of $g(\rho)$ which has p rows and r columns. Then using the differentiation rules of Section 2,

$$(\partial/\partial\rho')f[x,g(\rho)] = (\partial/\partial\theta')f[x,g(\rho)]G(\rho)$$

$$(\partial/\partial\rho')f[g(\rho)] = F[g(\rho)]G(\rho)$$

These facts can be used as a labor saving device when writing code for nonlinar optimization as seen in the examples.

EXAMPLE 1 (continued). Recalling that the response function is

$$f(x,\theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3},$$

reconsider the first hypothesis

$$H: \theta_1^o = 0.$$

This is an assertion that the data follows the model

$$y_t = \theta_2 x_{2t} + \theta_4 e^{\theta_3 x_{3t}} + e_t.$$

Fitting this model to the data of Table 1 by the modified Gauss-Newton method we have

Figure 9a. Illustration of Likelihood Ratio Test Computations with Example 1.

SAS Statements:

```
PROC NLIN DATA=EXAMPLE1 METHOD=GAUSS ITER=50 CONVERGENCE=1.0E-13;
PARMS  T2=1.01567967 T3=-1.11569714 T4=-0.50490286;  T1=0;
MODEL Y=T1*X1+T2*X2+T4*EXP(T3*X3);
DER.T2=X2; DER.T3=T4*X3*EXP(T3*X3); DER.T4=EXP(T3*X3);
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

NON-LINEAR LEAST SQUARES ITERATIVE PHASE

DEPENDENT VARIABLE: Y METHOD: GAUSS-NEWTON

ITERATION	T2	T3	T4	RESIDUAL SS
0	1.01567967	-1.11569714	-0.50490286	0.04054968
1	1.00289158	-1.14446980	-0.51206647	0.03543349
2	1.00297335	-1.14082057	-0.51178607	0.03543299
3	1.00296493	-1.14128672	-0.51182738	0.03543298
4	1.00296604	-1.14122778	-0.51182219	0.03543298
5	1.00296590	-1.14123524	-0.51182285	0.03543298
6	1.00296592	-1.14123430	-0.51182276	0.03543298
7	1.00296592	-1.14123442	-0.51182277	0.03543298

NOTE: CONVERGENCE CRITERION MET.

S T A T I S T I C A L A N A L Y S I S S Y S T E M 2

NON-LINEAR LEAST SQUARES SUMMARY STATISTICS DEPENDENT VARIABLE Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	3	26.34100467	8.78033489
RESIDUAL	27	0.03543298	0.00131233
UNCORRECTED TOTAL	30	26.37643764	
(CORRECTED TOTAL)	29	0.71895291	

PARAMETER	ESTIMATE	ASYMPTOTIC STD. ERROR	ASYMPTOTIC 95 % CONFIDENCE INTERVAL LOWER	UPPER
T2	1.00296592	0.00813053	0.98628359	1.01964825
T3	-1.14123442	0.17446900	-1.49921245	-0.78325638
T4	-0.51182277	0.02718622	-0.56760385	-0.45604169

ASYMPTOTIC CORRELATION MATRIX OF THE PARAMETERS

	T2	T3	T4
T2	1.000000	0.400991	-0.120866
T3	0.400991	1.000000	0.565235
T4	-0.120866	0.565235	1.000000

$$SSE_{reduced} = 0.03543298 \qquad \text{(from Figure 9a)}$$

Previously we computed

$$SSE_{full} = 0.03049554 \qquad \text{(from Figure 5a).}$$

The likelihood ratio statistic is

$$L = \frac{(SSE_{reduced} - SSE_{full})/q}{(SSE_{reduced})/(n - p)}$$

$$= \frac{(0.03543298 - 0.03049554)/1}{0.03049554/26}$$

$$= 4.210.$$

Comparing with the critical point

$$F^{-1}(.95;\ 1,\ 26) = 4.22$$

one fails to reject the null hypothesis at the 95% level.

Reconsider the second hypothesis

$$H:\ \theta_3\theta_4 e^{\theta_3} = 1/5$$

which can be rewritten as

$$H:\ \theta_4 = 1/(5\theta_3 e^{\theta_3}).$$

Then writing

$$g(\rho) = \begin{pmatrix} \rho_1 \\ \rho_2 \\ 1/(5\rho_3 e^{\rho_3}) \end{pmatrix}$$

an equivalent form of the null hypothesis is that

$$H: \theta^o = g(\rho) \text{ for some } \rho^o.$$

One can fit the null model in one of two ways. The first, fit directly the model

$$y_t = \rho_1 x_{1t} + \rho_2 x_{2t} + (5\rho_3)^{-1} e^{\rho_3 (x_{3t}-1)} + e_t.$$

The second,

1. Given ρ, set $\theta = g(\rho)$.

2. Use the code written previously (Figure 5a) to compute $f(x,\theta)$ and $(\partial/\partial\theta')f(x,\theta)$ given θ.

3. Use

$$(\partial/\partial\rho')f[x,g(\rho)] = \{(\partial/\partial\theta')f[x,g(\rho)]\}G(\rho)$$

to compute the partial derivatives with respect to ρ; recall that

$$G(\rho) = (\partial/\partial\rho')g(\rho).$$

We use this second method to fit the reduced model in Figure 9b. We have

$$G(\rho) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -(5\rho_3 e^{\rho_3})^{-2}(5e^{\rho_3} + 5\rho_3 e^{\rho_3}) \end{pmatrix}$$

If

$$(\partial/\partial\theta')f(x,\theta) = (DER\_T1,\ DER\_T2,\ DER\_T3,\ DER\_T4)$$

then to compute

$$(\partial/\partial\rho')f[x,g(\rho)] = (DER.R1,\ DER.R2,\ DER.R3)$$

one codes

```
DER.R1  =  DER_T1
DER.R2  =  DER_T2
DER.R3  =  DER_T3 + DER_T4 * (-T4**2) * (5*EXP(R3) + 5*R3*EXP(R3))
```

where

```
T4  =  1/(5*R3*EXP(R3))
```

Figure 9b. Illustration of Likelihood Ratio Test Computations with Example 1.

SAS Statements:

```
PROC NLIN DATA=EXAMPLE1 METHOD=GAUSS ITER=60 CONVERGENCE=1.0E-8;
PARMS R1=-0.02588970 R2=1.01567967 R3=-1.11569714;
T1=R1; T2=R2; T3=R3; T4=1/(5*R3*EXP(R3));
MODEL Y=T1*X1+T2*X2+T4*EXP(T3*X3);
DER_T1=X1; DER_T2=X2; DER_T3=T4*X3*EXP(T3*X3); DER_T4=EXP(T3*X3);
DER.R1=DER_T1; DER.R2=DER_T2;
DER.R3=DER_T3+DER_T4*(-T4**2)*(5*EXP(R3)+5*R3*EXP(R3));
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

NON-LINEAR LEAST SQUARES ITERATIVE PHASE

DEPENDENT VARIABLE: Y METHOD: GAUSS-NEWTON

ITERATION	R1	R2	R3	RESIDUAL SS
0	-0.02588970	1.01567967	-1.11569714	0.03644046
1	-0.02286308	1.01860305	-1.19237581	0.03502362
2	-0.02314184	1.02019397	-1.13249955	0.03500414
3	-0.02291862	1.01903284	-1.18159656	0.03497186
4	-0.02309964	1.02003652	-1.14220257	0.03496229
5	-0.02295240	1.01926378	-1.17465123	0.03495011
6	-0.02307276	1.01992190	-1.14831568	0.03494536
7	-0.02297427	1.01940189	-1.17003037	0.03494040
8	-0.02305506	1.01984017	-1.15230734	0.03493808
9	-0.02298878	1.01948877	-1.16691829	0.03493597
10	-0.02304322	1.01978274	-1.15495732	0.03493486
11	-0.02299850	1.01954494	-1.16481311	0.03493394
12	-0.02303525	1.01974282	-1.15673110	0.03493341
13	-0.02300504	1.01958186	-1.16338723	0.03493301
14	-0.02302988	1.01971531	-1.15792350	0.03493276
15	-0.02300946	1.01960636	-1.16242136	0.03493258
16	-0.02302625	1.01969645	-1.15872697	0.03493246
17	-0.02301245	1.01962272	-1.16176727	0.03493238
18	-0.02302380	1.01968358	-1.15926909	0.03493233
19	-0.02301447	1.01963370	-1.16132448	0.03493229
20	-0.02302214	1.01967482	-1.15963516	0.03493227
21	-0.02301583	1.01964108	-1.16102482	0.03493225
22	-0.02302102	1.01966888	-1.15988247	0.03493224
23	-0.02301675	1.01964605	-1.16082207	0.03493223
24	-0.02302026	1.01966484	-1.16004961	0.03493223
25	-0.02301738	1.01964941	-1.16068492	0.03493222
26	-0.02301975	1.01966211	-1.16016258	0.03493222
27	-0.02301780	1.01965167	-1.16059216	0.03493222
28	-0.02301940	1.01966026	-1.16023895	0.03493222
29	-0.02301808	1.01965320	-1.16052942	0.03493222
30	-0.02301917	1.01965901	-1.16029058	0.03493222
31	-0.02301828	1.01965423	-1.16048699	0.03493222

NOTE: CONVERGENCE CRITERION MET.

S T A T I S T I C A L A N A L Y S I S S Y S T E M 2

NON-LINEAR LEAST SQUARES SUMMARY STATISTICS DEPENDENT VARIABLE Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	3	26.34150543	8.78050181
RESIDUAL	27	0.03493222	0.00129379
UNCORRECTED TOTAL	30	26.37643764	
(CORRECTED TOTAL)	29	0.71895291	

PARAMETER	ESTIMATE	ASYMPTOTIC STD. ERROR	ASYMPTOTIC 95 % CONFIDENCE INTERVAL LOWER	UPPER
R1	-0.02301828	0.01315496	-0.05000981	0.00397326
R2	1.01965423	0.01009676	0.99893755	1.04037092
R3	-1.16048699	0.16302087	-1.49497559	-0.82599838

as shown in Figure 9b.

We have

$$SSE_{reduced} = 0.03493222 \qquad \text{(from Figure 9b)}$$

$$SSE_{full} = 0.03049554 \qquad \text{(from Figure 5a)}$$

$$L = \frac{(0.03493222 - 0.03049554)/1}{0.03049554/26}$$

$$= 3.783.$$

As $F^{-1}(.95;\ 1,\ 26) = 4.22$ one fails to reject the null hypothesis at the 5% level.

Reconsidering the third hypothesis

$$H:\ \theta_1 = 0 \quad \text{and} \quad \theta_3\theta_4 e^{\theta_3} = 0$$

which may be rewritten as

$$H:\ \theta^o = g(\rho) \text{ for some } \rho^o$$

with

$$g(\rho) = \begin{pmatrix} 0 \\ \rho_2 \\ \rho_3 \\ 1/(5\rho_3 e^{\rho_3}) \end{pmatrix}$$

we have

Figure 9c. Illustration of Likelihood Ratio Test Computations with Example 1.

SAS Statements:

```
PROC NLIN DATA=EXAMPLE1 METHOD=GAUSS ITER=60 CONVERGENCE=1.0E-8;
PARMS R2=1.01965423 R3=-1.16048699;  R1=0;
T1=R1; T2=R2; T3=R3; T4=1/(5*R3*EXP(R3));
MODEL Y=T1*X1+T2*X2+T4*EXP(T3*X3);
DER_T1=X1; DER_T2=X2; DER_T3=T4*X3*EXP(T3*X3); DER_T4=EXP(T3*X3);
DER.R2=DER_T2; DER.R3=DER_T3+DER_T4*(-T4**2)*(5*EXP(R3)+5*R3*EXP(R3));
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

NON-LINEAR LEAST SQUARES ITERATIVE PHASE

DEPENDENT VARIABLE: Y METHOD: GAUSS-NEWTON

ITERATION	R2	R3	RESIDUAL SS
0	1.01965423	-1.16048699	0.04287983
1	1.00779498	-1.17638081	0.03890362
2	1.00807441	-1.16332560	0.03890234
3	1.00784845	-1.17411590	0.03890127
4	1.00803764	-1.16523771	0.03890066
5	1.00788362	-1.17257272	0.03890018
6	1.00801199	-1.16653150	0.03889989
7	1.00790702	-1.17152084	0.03889967
8	1.00799423	-1.16740905	0.03889954
9	1.00792271	-1.17080393	0.03889944
10	1.00798200	-1.16800508	0.03889937
11	1.00793329	-1.17031543	0.03889933
12	1.00797361	-1.16841024	0.03889930
13	1.00794043	-1.16998265	0.03889928
14	1.00796787	-1.16868578	0.03889926
15	1.00794527	-1.16975601	0.03889925
16	1.00796394	-1.16887322	0.03889925
17	1.00794856	-1.16960168	0.03889924
18	1.00796126	-1.16900077	0.03889924
19	1.00795079	-1.16949660	0.03889924
20	1.00795944	-1.16908756	0.03889923
21	1.00795231	-1.16942506	0.03889923
22	1.00795819	-1.16914663	0.03889923
23	1.00795334	-1.16937636	0.03889923
24	1.00795735	-1.16918683	0.03889923

NOTE: CONVERGENCE CRITERION MET.

S T A T I S T I C A L A N A L Y S I S S Y S T E M 2

NON-LINEAR LEAST SQUARES SUMMARY STATISTICS DEPENDENT VARIABLE Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	2	26.33753841	13.16876921
RESIDUAL	28	0.03889923	0.00138926
UNCORRECTED TOTAL	30	26.37643764	
(CORRECTED TOTAL)	29	0.71895291	

PARAMETER	ESTIMATE	ASYMPTOTIC STD. ERROR	ASYMPTOTIC 95 % CONFIDENCE INTERVAL LOWER	UPPER
R2	1.00795735	0.00769931	0.99218613	1.02372856
R3	-1.16918683	0.17039162	-1.51821559	-0.82015808

ASYMPTOTIC CORRELATION MATRIX OF THE PARAMETERS

	R2	R3
R2	1.000000	0.467769
R3	0.467769	1.000000

$SSE_{reduced} = 0.03889923$ (from Figure 9c)

$SSE_{full} = 0.03049554$ (from Figure 5a)

$$L = \frac{(SSE_{reduced} - SSE_{full})/(p - r)}{(SSE_{full})/(n - p)}$$

$$= \frac{(0.03889923 - 0.03049554)/(4 - 2)}{(0.03049554)/(30 - 4)}$$

$$= 3.582.$$

Since $F^{-1}(.95; 2, 26) = 3.37$ one rejects the null hypothesis at the 5% level. I

It is not always easy to convert a parametric restriction $h(\theta) = 0$ to a functional dependency $\theta = g(\rho)$ analytically. However, all that is needed is the value of θ for given ρ and the value of $(\partial/\partial\rho)g(\rho)$ for given ρ. This allows substitution of numerical methods for analytical methods in the determination of $g(\rho)$. We illustrate with the next example.

EXAMPLE 2 (continued). Recall that the amount of substance in compartment B at time x is given by the response function

$$f(x,\theta) = \theta_1(e^{-x\theta_2} - e^{-x\theta_1})/(\theta_1 - \theta_2).$$

By differentiating with respect to x and setting the derivative to zero one has that the time at which the maximum amount of substance present in compartment B is

$$\hat{x} = (\ell n\theta_1 - \ell n\theta_2)/(\theta_1 - \theta_2).$$

The unconstrained fit of this model is shown in Figure 10a. Suppose that we want to test

$$H: \hat{x} = 1 \quad \text{against} \quad A: \hat{x} \neq 1.$$

This requires that

$$h(\theta) = (\ell n\theta_1 - \ell n\theta_2)/(\theta_1 - \theta_2) - 1$$

be converted to a functional dependency if one is to be able to use unconstrained optimization methods. To do this numerically, set $\theta_2 = \rho$. Then the problem is to solve the equation

$$\theta_1 = \ell n\theta_1 + \rho - \ell n\rho$$

for θ_1. Stated differently, we are trying to find a fixed point of the equation

$$z = \ell nz + \text{const.}$$

But ℓnz + const. is a contraction mapping for $z > 1$--the derivative with respect to z is less than one--so that a fixed point can be found by successive substitution

Figure 10a. Illustration of Likelihood Ratio Test Computations with Example 2.

SAS Statements:

```
PROC NLIN DATA=EG2B METHOD=GAUSS ITER=50 CONVERGENCE=1.E-10;
PARMS T1=1.4 T2=.4;
MODEL Y=T1*(EXP(-T2*X)-EXP(-T1*X))/(T1-T2);
DER.T1=-T2*(EXP(-T2*X)-EXP(-T1*X))/(T1-T2)**2+T1*X*EXP(-T1*X)/(T1-T2);
DER.T2=T1*(EXP(-T2*X)-EXP(-T1*X))/(T1-T2)**2-T1*X*EXP(-T2*X)/(T1-T2);
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

NON-LINEAR LEAST SQUARES ITERATIVE PHASE

DEPENDENT VARIABLE: Y METHOD: GAUSS-NEWTON

ITERATION	T1	T2	RESIDUAL SS
0	1.40000000	0.40000000	0.00567248
1	1.37373983	0.40266678	0.00545775
2	1.37396974	0.40265518	0.00545774
3	1.37396966	0.40265518	0.00545774

NOTE: CONVERGENCE CRITERION MET.

S T A T I S T I C A L A N A L Y S I S S Y S T E M 2

NON-LINEAR LEAST SQUARES SUMMARY STATISTICS DEPENDENT VARIABLE Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	2	2.68129496	1.34064748
RESIDUAL	10	0.00545774	0.00054577
UNCORRECTED TOTAL	12	2.68675270	
(CORRECTED TOTAL)	11	0.21359486	

PARAMETER	ESTIMATE	ASYMPTOTIC STD. ERROR	ASYMPTOTIC 95 % CONFIDENCE INTERVAL LOWER	UPPER
T1	1.37396966	0.04864622	1.26557844	1.48236088
T2	0.40265518	0.01324390	0.37314574	0.43216461

ASYMPTOTIC CORRELATION MATRIX OF THE PARAMETERS

	T1	T2
T1	1.000000	0.236174
T2	0.236174	1.000000

Figure 10b. Illustration of Likelihood Ratio Test Computations with Example 2.

SAS Statements:

```
PROC NLIN DATA=EG2B METHOD=GAUSS ITER=50 CONVERGENCE=1.E-10;
PARMS RHO=.40265518;
T2=RHO;
Z1=1.4; Z2=0; C=T2-LOG(T2);
L1: IF ABS(Z1-Z2)>1.E-13 THEN DO; Z2=Z1; Z1=LOG(Z1)+C; GO TO L1; END;
T1=Z1;
NU2=T1*(EXP(-T2*X)-EXP(-T1*X))/(T1-T2);
DER_T1=-T2*(EXP(-T2*X)-EXP(-T1*X))/(T1-T2)**2+T1*X*EXP(-T1*X)/(T1-T2);
DER_T2=T1*(EXP(-T2*X)-EXP(-T1*X))/(T1-T2)**2-T1*X*EXP(-T2*X)/(T1-T2);
DER_RHO=DER_T1*(1-1/T2)/(1-1/T1)+DER_T2;
MODEL Y=NU2; DER.RHO=DER_RHO;
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

NON-LINEAR LEAST SQUARES ITERATIVE PHASE

DEPENDENT VARIABLE: Y METHOD: GAUSS-NEWTON

ITERATION	RHO	RESIDUAL SS
0	0.40265518	0.07004386
1	0.46811176	0.04654328
2	0.47688375	0.04621215
3	0.47750162	0.04621056
4	0.47754034	0.04621055
5	0.47754274	0.04621055
6	0.47754289	0.04621055

NOTE: CONVERGENCE CRITERION MET.

S T A T I S T I C A L A N A L Y S I S S Y S T E M 2

NON-LINEAR LEAST SQUARES SUMMARY STATISTICS DEPENDENT VARIABLE Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	1	2.64054214	2.64054214
RESIDUAL	11	0.04621055	0.00420096
UNCORRECTED TOTAL	12	2.68675270	
(CORRECTED TOTAL)	11	0.21359486	

PARAMETER	ESTIMATE	ASYMPTOTIC STD. ERROR	ASYMPTOTIC 95 % CONFIDENCE INTERVAL LOWER	UPPER
RHO	0.47754289	0.03274044	0.40548138	0.54960439

ASYMPTOTIC CORRELATION MATRIX OF THE PARAMETERS

	RHO
RHO	1.000000

$$z_1 = \ell n z_0 + \text{const.}$$

$$z_2 = \ell n z_1 + \text{const.}$$

$$\vdots$$

$$z_{i+1} = \ell n z_i + \text{const.}$$

$$\vdots$$

This sequence $\{z_{i+1}\}$ will converge to the fixed point.

To compute $(\partial/\partial\rho)g(\rho)$ we apply the implicit function theorem to

$$\theta_1(\rho) - \ell n[\theta_1(\rho)] = \rho - \ell n\rho$$

We have

$$(\partial/\partial\theta_1)\{\theta_1(\rho) - \ell n[\theta_1(\rho)]\}\ (\partial/\partial\rho)\theta_1(\rho) = (\partial/\partial\rho)(\rho - \ell n\rho)$$

or

$$(\partial/\partial\rho)\theta_1(\rho) = (1 - 1/\rho)/[1 - 1/\theta_1(\rho)]$$

Then the Jacobian of $\theta = g(\rho)$ is

$$(\partial/\partial\rho') = \begin{pmatrix} (1 - 1/\rho)/[1 - 1/\theta_1(\rho)] \\ \\ 1 \end{pmatrix}$$

and

$$(\partial/\partial\rho)f[x,g(\rho)] = \{[(\partial/\partial\theta_1)f(x,\theta)](1-1/\rho)/(1-1/\theta_1) + (\partial/\partial\theta_2)f(x,\theta_2)\}\Big|_{\theta=g(\rho)}$$

These ideas are illustrated in Figure 10b. We have

$$SSE_{full} = 0.00545774 \qquad \text{(from Figure 10a)}$$

$$SSE_{reduced} = 0.04621055 \qquad \text{(from Figure 10b)}$$

$$L = \frac{(SSE_{reduced} - SSE_{full})/q}{(0.00545774)/(12 - 2)}$$

$$= 74.670.$$

As $F^{-1}(.95;\ 1,\ 10) = 4.96$ one rejects H. ∎

Now let us turn our attention to the computation of the power of the likelihood ratio test. That is, for data that follow the model

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

$$e_t \text{ iid. } n(0,\sigma^2),$$

$$t = 1,\ 2,\ \ldots,\ n,$$

we should like to compute

$$P(L > L_\alpha |\ \theta^o,\ \sigma^2,\ n),$$

the probability that the likelihood ratio test rejects at level α given θ^o,

σ^2, and n where $F_\alpha = F^{-1}(1 - \alpha;\ q,\ n-p)$. To do this, note that the test that rejects when

$$L = \frac{(SSE_{reduced} - SSE_{full})/q}{(SSE_{full})/(n - p)} > F_\alpha$$

is equivalent to the test that rejects when

$$(SSE_{reduced})/(SSE_{full}) > c$$

where

$$c_\alpha = 1 + qF_\alpha/(n - p).$$

In Chapter 4 we shall show that

$$n/(SSE_{full}) = n/e' P_F^{\perp} e + o_p(1/n)$$

where

$$P_F^{\perp} = I - F(F'F)^{-1}F';$$

Recall that $F = (\partial/\partial\theta)f(\theta^o)$. Then it remains to obtain an approximation to $(SSE_{reduced})/n$ in order to approximate $(SSE_{reduced})/(SSE_{full})$. To this end, let

$$\theta_n^* = g(\rho_n^o)$$

where

$$\rho_n^o \text{ minimizes } \sum_{t=1}^{n} \{f(x_t,\theta^o) - f[x_t,g(\rho)]\}^2.$$

Recall that $g(\rho)$ is the mapping from R^r into R^p that describes the null hypothesis--H: $\theta^o = g(\rho)$ for some ρ^o; $r = p-q$. The point θ_n^* may be interpreted as that point which is being estimated by the constrained estimator $\tilde{\theta}_n$ in the sense that $\sqrt{n}(\tilde{\theta}_n - \theta_n^*)$ converges in distribution to the multivariate normal distribution; see Chapter 3 for details. Under this interpretation,

$$\delta = f(\theta^o) - f(\theta_n^*)$$

may be interepreted as the prediction bias. We shall show later (Chapter 4) that what one's intuition would suggest is true.[1]

$$(SSE_{reduced})/n = (e + \delta)' P_{FG}^{\perp}(e + \delta)/n + o_p(1/n)$$

where

$$P_{FG}^{\perp} = I - FG(G'F'FG)^{-1}G'F',$$

$$F = (\partial/\partial\theta')f(\theta^o),$$

---

[1]One's intuition might also suggest that the Jacobian $F(\theta) = (\partial/\partial\theta')f(\theta)$ ought to be evaluated at θ_n^* rather than θ^o, especially in view of Theorems 6 and 13 of Chapter 3. This is correct, the discrepancy caused by the substitution of θ^o for θ_n^* has been absorbed into the $o_p(1/n)$ term in order to permit the derivation of the small sample distribution of the random variable X. Details are in Chapter 4.

$$G = (\partial/\partial\rho')g(\rho_n^o).$$

It follows from the characterizations of the residual sum of squares for the full and reduced models that

$$(SSE_{reduced})/(SSE_{full}) = X + o_p(1/n)$$

where

$$X = (e + \delta)'P_{FG}^{\perp}(e + \delta)/e'P_F^{\perp}e.$$

The idea, then, is to approximate the probability $P(L > F_\alpha |\ \theta^o, \sigma^2, n)$ by the probability $P(X > c_\alpha |\ \theta^o, \sigma^2, n)$. The distribution function of the random variable X is for $x > 1$ (Problem 4).

$$H(x;\ \nu_1, \nu_2, \lambda_1, \lambda_2)$$

$$= 1 - \int_0^\infty G[t/(x-1) + 2x\lambda_2/(x-1)^2;\ \nu_2,\ \lambda_2/(x-1)^2]g(t;\ \nu_1,\ \lambda_1)dt$$

where $q(t;\nu,\lambda)$ denotes the non-central chi-square density function with ν degress of freedom and non-centrality parameter λ and $G(t;\nu,\lambda)$ denotes the corresponding distribution function (Appendix 1). The two degrees of freedom entries are

$$\nu_1 = q = p - r$$

Table 6. Power of the Likelihood Ratio Test at the 5% Level.

λ_2	λ_1										
	0	.5	1	2	3	4	5	6	8	10	12
					A. $\nu_1=1$, $\nu_2=10$						
0.0	.050	.148	.249	.440	.599	.722	.813	.876	.949	.980	.992
.0001	.050	.148	.249	.440	.599	.722	.813	.876	.949	.980	.992
.001	.050	.148	.249	.440	.599	.723	.813	.876	.949	.980	.992
.01	.051	.150	.251	.442	.601	.724	.814	.877	.949	.980	.992
.1	.063	.168	.272	.462	.617	.735	.821	.882	.951	.980	.992
					B. $\nu_1=1$, $\nu_2=20$						
0.0	.050	.159	.271	.478	.645	.768	.853	.909	.967	.989	.996
.0001	.050	.159	.271	.478	.645	.768	.853	.909	.967	.989	.996
.001	.050	.159	.271	.478	.645	.768	.853	.909	.967	.989	.996
.01	.051	.161	.273	.480	.647	.769	.853	.909	.967	.989	.996
.1	.065	.181	.296	.501	.663	.780	.860	.913	.968	.989	.996
					C. $\nu_1=1$, $\nu_2=30$						
0.0	.050	.163	.278	.490	.659	.781	.864	.917	.972	.991	.997
.0001	.050	.163	.278	.490	.659	.781	.864	.917	.972	.991	.997
.001	.050	.163	.278	.491	.659	.781	.864	.917	.972	.991	.997
.01	.051	.165	.280	.493	.661	.782	.864	.918	.972	.991	.997
.1	.065	.185	.303	.514	.676	.792	.871	.921	.973	.991	.997
					D. $\nu_1=2$, $\nu_2=10$						
0.0	.050	.111	.178	.318	.454	.575	.677	.759	.873	.936	.969
.0001	.050	.111	.178	.318	.454	.575	.677	.759	.873	.936	.969
.001	.050	.111	.178	.318	.454	.575	.677	.759	.873	.936	.969
.01	.051	.112	.179	.320	.456	.576	.678	.760	.873	.936	.969
.1	.058	.122	.192	.334	.469	.588	.688	.767	.877	.938	.970
					E. $\nu_1=2$, $\nu_2=20$						
0.0	.050	.121	.199	.364	.517	.647	.749	.827	.922	.968	.987
.0001	.050	.121	.199	.364	.517	.647	.749	.827	.922	.968	.987
.001	.050	.121	.200	.364	.517	.647	.750	.827	.922	.968	.987
.01	.051	.122	.201	.365	.519	.648	.750	.828	.923	.968	.987
.1	.060	.135	.216	.382	.534	.660	.759	.834	.925	.969	.987
					F. $\nu_1=2$, $\nu_2=30$						
0.0	.050	.124	.208	.381	.539	.671	.773	.847	.936	.975	.991
.0001	.050	.124	.208	.381	.539	.671	.773	.847	.936	.975	.991
.001	.050	.125	.208	.381	.539	.671	.773	.847	.936	.975	.991
.01	.051	.126	.210	.382	.541	.672	.774	.848	.936	.975	.991
.1	.060	.139	.226	.400	.556	.684	.782	.854	.938	.976	.991

Continued Next Page

Table 6. Continued.

λ_2	λ_1										
	0	.5	1	2	3	4	5	6	8	10	12
					G. $\nu_1=3$, $\nu_2=10$						
0.0	.050	.094	.145	.255	.368	.477	.576	.662	.794	.881	.933
.0001	.050	.094	.145	.255	.368	.477	.576	.662	.794	.881	.933
.001	.050	.095	.145	.255	.368	.477	.576	.662	.794	.881	.933
.01	.051	.095	.146	.256	.369	.478	.577	.662	.795	.881	.934
.1	.056	.103	.155	.267	.381	.489	.586	.670	.800	.884	.935
					H. $\nu_1=3$, $\nu_2=20$						
0.0	.050	.104	.165	.300	.436	.561	.668	.755	.874	.940	.973
.0001	.050	.104	.165	.300	.436	.561	.668	.755	.874	.940	.973
.001	.050	.104	.165	.300	.437	.561	.668	.755	.874	.940	.973
.01	.051	.105	.166	.302	.438	.562	.669	.755	.875	.940	.973
.1	.057	.114	.178	.316	.452	.574	.679	.763	.878	.942	.973
					I. $\nu_1=3$, $\nu_2=30$						
0.0	.050	.107	.173	.318	.462	.591	.699	.785	.897	.954	.981
.0001	.050	.107	.173	.318	.462	.591	.699	.785	.897	.954	.981
.001	.050	.107	.173	.318	.462	.592	.699	.785	.897	.954	.981
.01	.051	.108	.175	.320	.464	.593	.700	.785	.897	.954	.981
.1	.058	.119	.187	.335	.478	.605	.710	.792	.900	.956	.981

$$\nu_2 = n - p$$

and the non-centrality parameters are

$$\lambda_1 = \delta'(P_F - P_{FG})\delta/(2\sigma^2)$$

$$\lambda_2 = \delta' P_F^{\perp}\delta/(2\sigma^2)$$

where $P_F = F(F'F)^{-1}F'$, $P_{FG} = FG(G'F'FG)^{-1}G'F'$, and $P_F^{\perp} = I - P_F$. This distribution is partially tabulated in Table 6. Let us illustrate the computations necessary to use these tables and check the accuracy of the approximation of $P(L > F_\alpha)$ by $P(X > c_\alpha)$ by Monte Carlo simulation using Example 1.

EXAMPLE 1 (continued). Recalling that

$$f(x,\theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}$$

let us approximate the probability that the likelihood ratio test rejects the following three hypotheses at the 5% level when the true values of the parameters are

$$\theta^o = (.03,\ 1,\ -1.4,\ -.5)',$$

$$\sigma^2 = .001.$$

The three null hypotheses are:

$$H_1: \theta_1 = 0,$$

$$H_2: \theta_3\theta_4 e^{\theta_3} = 1/5$$

$$H_3: \theta_3 = 0 \quad \text{and} \quad \theta_3\theta_4 e^{\theta_3} = 1/5$$

The computational chore is to compute for each hypothesis:

$$\rho_n^o \text{ minimizing } \sum_{t=1}^{n} \{f(x_t,\theta^o) - f[x_t,g(\rho)]\}$$

$$\delta = f(\theta^o) - f(\theta_n^*), \quad \theta_n^* = g(\rho_n^o)$$

$$\delta' P_F \delta, \quad \delta' P_{FG} \delta, \quad \text{and} \quad \delta'\delta.$$

With these, the non-centrality parameters

$$\lambda_1 = (\delta' P_F \delta - \delta' P_{FG} \delta)/(2\sigma^2)$$

$$\lambda_2 = (\delta'\delta - \delta P_F \delta)/(2\sigma^2)$$

are easily computed. As usual, there are a variety of strategies that one might employ.

To compute δ, the easiest approach is to notice that minimizing

$$\sum_{t=1}^{n} \{f(x_t,\theta^o) - f[x_t,g(\rho)]\}$$

is no different than minimizing

$$\sum_{t=1}^{n} \{y_t - f[x_t, g(\rho)]\}.$$

One simply replaces y_t by $f(x_t, \theta^o)$ and uses the modified Gauss-Newton method, the Levenberg-Marquardt method, or whatever.

To compute $\delta' P_F \delta$ one can either proceed directly using a programming language such as PROC MATRIX or make the following observation. If one regresses δ on F with no intercept term using a linear regression procedure then the analysis of variance table printed by the program will have the following entries

Source	d.f.	Sum of Squares
Regression	p	$\delta' F(F'F)^{-1}F'\delta$
Error	n - p	$\delta'\delta - \delta' F(F'F)^{-1}F'\delta$
Total	n	$\delta'\delta$

One can just read off

$$\delta' P_F \delta = \delta' F(F'F)F'\delta$$

from the analysis of variance table. Similarly for a regression of δ on FG.

Figures 11a, 11b, and 11c illustrate these ideas for the hypothesis H_1, H_2, and H_3.

For the first hypothesis we have

$$\delta'\delta = 0.006668583 \qquad \text{(from Figure 11a)}$$

$$\delta'P_F\delta = 0.006668583 \qquad \text{(from Figure 11a)}$$

$$\delta'P_{FG}\delta = 3.25 \times 10^{-9} \qquad \text{(from Figure 11a)}$$

whence

$$\lambda_1 = (\delta'P_F\delta - \delta'P_{FG}\delta)/(2\sigma^2)$$
$$= (0.006668583 - 3.25 \times 10^{-9})/(2 \times .001)$$
$$= 3.3343$$

$$\lambda_2 = (\delta'\delta - \delta'P_F\delta)/(2\sigma^2)$$
$$= (0.006668583 - 0.006668583)/(2 \times .001)$$
$$= 0$$

$$c_\alpha = 1 + qF_\alpha/(n - p)$$
$$= 1 + (1)(4.22)/26$$
$$= 1.1623$$

Computing $1 - H(1.1623;\ 1,\ 26,\ \lambda_1,\ \lambda_2)$ by interpolating from Table 6 we obtain

$$P(X > c_\alpha) = .700$$

as an approximation to $P(L > F_\alpha)$. Later we shall show that tables of the non-central F will usually be accurate enough so there is usually no need for special tables or special computations.

Figure 11a. Illustration of Likelihood Ratio Test Power Computations with Example 1.

SAS Statements:

```
DATA WORK01;  SET EXAMPLE1;  T1=.03;  T2=1;  T3=-1.4;  T4=-.5;
YDUMMY=T1*X1+T2*X2+T4*EXP(T3*X3);
F1=X1;  F2=X2;  F3=T4*X3*EXP(T3*X3);  F4=EXP(T3*X3);
DROP T1 T2 T3 T4;
PROC NLIN DATA=WORK01 METHOD=GAUSS ITER=50 CONVERGENCE=1.0E-13;
PARMS T2=1 T3=-1.4 T4=-.5;  T1=0;
MODEL YDUMMY=T1*X1+T2*X2+T4*EXP(T3*X3);
DER.T2=X2; DER.T3=T4*X3*EXP(T3*X3); DER.T4=EXP(T3*X3);
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

NON-LINEAR LEAST SQUARES ITERATIVE PHASE

DEPENDENT VARIABLE: YDUMMY METHOD: GAUSS-NEWTON

ITERATION	T2	T3	T4	RESIDUAL SS
0	1.00000000	-1.40000000	-0.50000000	0.01350000
1	1.01422090	-1.39717572	-0.49393589	0.00666859
2	1.01422435	-1.39683401	-0.49391057	0.00666858
3	1.01422476	-1.39679638	-0.49390747	0.00666858
4	1.01422481	-1.39679223	-0.49390713	0.00666858
5	1.01422481	-1.39679178	-0.49390709	0.00666858
6	1.01422481	-1.39679173	-0.49390708	0.00666858

NOTE: CONVERGENCE CRITERION MET.

SAS Statements:

```
DATA WORK02;  SET WORK01;
T1=0;  T2=1.01422481;  T3=-1.39679173;  T4=-0.49390708;
DELTA=YDUMMY-(T1*X1+T2*X2+T4*EXP(T3*X3));
FG1=F2;  FG2=F3;  FG3=F4;  DROP T1 T2 T3 T4;
PROC REG DATA=WORK02;  MODEL DELTA=F1 F2 F3 F4 / NOINT;
PROC REG DATA=WORK02;  MODEL DELTA=FG1 FG2 FG3 / NOINT;
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

DEP VARIABLE: DELTA

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	4	0.006668583	0.001667146	999999.990	0.0001
ERROR	26	2.89364E-13	1.11294E-14		
U TOTAL	30	0.006668583			

S T A T I S T I C A L A N A L Y S I S S Y S T E M 2

DEP VARIABLE: DELTA

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	3	3.25099E-09	1.08366E-09	0.000	1.0000
ERROR	27	0.00666858	0.0002469844		
U TOTAL	30	0.006668583			

For the second hypothesis we have

$$\delta'\delta = 0.01321589 \quad \text{(from Figure 11b)}$$

$$\delta'P_F\delta = 0.013215 \quad \text{(from Figure 11b)}$$

$$\delta'\delta - \delta'P_F\delta = 0.00000116542 \quad \text{(from Figure 11b)}$$

$$\delta'P_{FG}\delta = 0.0001894405 \quad \text{(from Figure 11b)}$$

whence

$$\lambda_1 = (\delta'P_F\delta - \delta'P_{FG}\delta)/(2\sigma^2)$$
$$= (0.013215 - 0.0001894405)/(2 \times .001)$$
$$= 6.5128$$

$$\lambda_2 = (\delta'\delta - \delta'P_F\delta)/(2 \times \sigma^2)$$
$$= (0.00000116542)/(2 \times .001)$$
$$= 0.0005827$$

$$c_\alpha = 1 + qF_\alpha/(n - p)$$
$$= 1 + (1)(4.22)/26$$
$$= 1.1623$$

Computing $1 - H(1.1623;\ 1,\ 26,\ \lambda_1,\ \lambda_2)$ as above we obtain

$$P(X > c_\alpha) = .935$$

as an approximation to $P(L > F_\alpha)$.

Figure 11b. Illustration of Likelihood Ratio Test Power Computations with Example 1.

SAS Statements:

```
DATA WORK01;  SET EXAMPLE1;  T1=.03;  T2=1;  T3=-1.4;  T4=-.5;
YDUMMY=T1*X1+T2*X2+T4*EXP(T3*X3);
F1=X1;  F2=X2;  F3=T4*X3*EXP(T3*X3);  F4=EXP(T3*X3);
DROP T1 T2 T3 T4;
PROC NLIN DATA=WORK01 METHOD=GAUSS ITER=50 CONVERGENCE=1.0E-13;
PARMS R1=.03 R2=1 R3=-1.4;
T1=R1; T2=R2; T3=R3; T4=1/(5*R3*EXP(R3));
MODEL YDUMMY=T1*X1+T2*X2+T4*EXP(T3*X3);
DER_T1=X1; DER_T2=X2; DER_T3=T4*X3*EXP(T3*X3); DER_T4=EXP(T3*X3);
DER.R1=DER_T1; DER.R2=DER_T2;
DER.R3=DER_T3+DER_T4*(-T4**2)*(5*EXP(R3)+5*R3*EXP(R3));
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

NON-LINEAR LEAST SQUARES ITERATIVE PHASE

DEPENDENT VARIABLE: YDUMMY METHOD: GAUSS-NEWTON

ITERATION	R1	R2	R3	RESIDUAL SS
0	0.03000000	1.00000000	-1.40000000	0.01867856
1	0.03363136	1.01008796	-1.12533963	0.01588546
2	0.03440842	1.00692167	-1.28648656	0.01344947
3	0.03425560	1.01002926	-1.25424342	0.01325389
4	0.03435915	1.00968253	-1.27776231	0.01321800
5	0.03433517	1.00977877	-1.27229450	0.01321601
6	0.03434071	1.00977225	-1.27354293	0.01321590
7	0.03433948	1.00978190	-1.27325579	0.01321589
8	0.03434008	1.00978565	-1.27338768	0.01321589
9	0.03433966	1.00978700	-1.27329144	0.01321589
10	0.03433976	1.00978669	-1.27331354	0.01321589
11	0.03433973	1.00978677	-1.27330847	0.01321589
12	0.03433974	1.00978675	-1.27330963	0.01321589
13	0.03433974	1.00978675	-1.27330936	0.01321589
14	0.03433974	1.00978675	-1.27330943	0.01321589

NOTE: CONVERGENCE CRITERION MET.

SAS Statements:

```
DATA WORK02;  SET WORK01;
R1=0.03433974;  R2=1.00978675;  R3=-1.27330943;
T1=R1; T2=R2; T3=R3; T4=1/(5*R3*EXP(R3));
DELTA=YDUMMY-(T1*X1+T2*X2+T4*EXP(T3*X3));
FG1=F1;  FG2=F2;  FG3=F3+F4*(-T4**2)*(5*EXP(R3)+5*R3*EXP(R3));
DROP T1 T2 T3 T4;
PROC REG DATA=WORK02;  MODEL DELTA=F1 F2 F3 F4 / NOINT;
PROC REG DATA=WORK02;  MODEL DELTA=FG1 FG2 FG3 / NOINT;
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

DEP VARIABLE: DELTA

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	4	0.013215	0.003303681	73703.561	0.0001
ERROR	26	.00000116542	4.48239E-08		
U TOTAL	30	0.013216			

S T A T I S T I C A L A N A L Y S I S S Y S T E M 2

DEP VARIABLE: DELTA

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	3	0.0001894405	.00006314682	0.131	0.9409
ERROR	27	0.013026	0.0004824611		
U TOTAL	30	0.013216			

Figure 11c. Illustration of Likelihood Ratio Test Power Computations with Example 1.

SAS Statements:

```
DATA WORK01;  SET EXAMPLE1;  T1=.03;  T2=1;  T3=-1.4;  T4=-.5;
YDUMMY=T1*X1+T2*X2+T4*EXP(T3*X3);
F1=X1;  F2=X2;  F3=T4*X3*EXP(T3*X3);  F4=EXP(T3*X3);
DROP T1 T2 T3 T4;
PROC NLIN DATA=WORK01 METHOD=GAUSS ITER=50 CONVERGENCE=1.0E-13;
PARMS R2=1 R3=-1.4;  R1=0;
T1=R1; T2=R2; T3=R3; T4=1/(5*R3*EXP(R3));
MODEL YDUMMY=T1*X1+T2*X2+T4*EXP(T3*X3);
DER_T1=X1; DER_T2=X2; DER_T3=T4*X3*EXP(T3*X3); DER_T4=EXP(T3*X3);
DER.R2=DER_T2;  DER.R3=DER_T3+DER_T4*(-T4**2)*(5*EXP(R3)+5*R3*EXP(R3));
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

NON-LINEAR LEAST SQUARES ITERATIVE PHASE

DEPENDENT VARIABLE: YDUMMY METHOD: GAUSS-NEWTON

ITERATION	R2	R3	RESIDUAL SS
0	1.00000000	-1.40000000	0.04431091
1	1.02698331	-1.10041642	0.02539361
2	1.02383184	-1.26840577	0.02235554
3	1.02719587	-1.25372059	0.02205576
4	1.02705467	-1.26454488	0.02204817
5	1.02709154	-1.26197184	0.02204774
6	1.02708616	-1.26258128	0.02204771
7	1.02708920	-1.2624367l	0.02204771
8	1.02708937	-1.26247100	0.02204771
9	1.02709018	-1.26245473	0.02204771
10	1.02709003	-1.26246672	0.02204771
11	1.02709006	-1.26246388	0.02204771
12	1.02709005	-1.26246455	0.02204771
13	1.02709006	-1.26246439	0.02204771

NOTE: CONVERGENCE CRITERION MET.

SAS Statements:

```
DATA WORK02;  SET WORK01;
R1=0;  R2=1.02709006;  R3=-1.26246439;
T1=R1; T2=R2; T3=R3; T4=1/(5*R3*EXP(R3));
DELTA=YDUMMY-(T1*X1+T2*X2+T4*EXP(T3*X3));
FG1=F2;  FG2=F3+F4*(-T4**2)*(5*EXP(R3)+5*R3*EXP(R3));
DROP T1 T2 T3 T4;
PROC REG DATA=WORK02;  MODEL DELTA=F1 F2 F3 F4 / NOINT;
PROC REG DATA=WORK02;  MODEL DELTA=FG1 FG2 / NOINT;
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

DEP VARIABLE: DELTA

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	4	0.022046	0.005511515	86947.729	0.0001
ERROR	26	.00000164811	6.33888E-08		
U TOTAL	30	0.022048			

S T A T I S T I C A L A N A L Y S I S S Y S T E M 2

DEP VARIABLE: DELTA

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	2	0.0001252535	.00006262677	0.080	0.9233
ERROR	28	0.021922	0.0007829449		
U TOTAL	30	0.022048			

For the third hypothesis we have

$\delta'\delta = 0.02204771$ (from Figure 11c)

$\delta'P_F\delta = 0.022046$ (from Figure 11c)

$\delta'\delta - \delta'P_F\delta = 0.00000164811$ (from Figure 11c)

$\delta'P_{FG}\delta = 0.0001252535$ (from Figure 11c)

whence

$$\lambda_1 = (\delta'P_F\delta - \delta'P_{FG}\delta)/(2\sigma^2)$$
$$= (0.022046 - 0.0001252535)/(2 \times .001)$$
$$= 10.9604$$

$$\lambda_2 = (\delta'\delta - \delta'P_F\delta)/(2 \times .001)$$
$$= (0.00000164811)/(2 \times .001)$$
$$= 0.0008241$$

$$c_\alpha = 1 + qF_\alpha/(n - p)$$
$$= 1 + (2)(3.37)/(26)$$
$$= 1.2592$$

Computing $1 - H(1.2592; 2, 26, \lambda_1, \lambda_2)$ as above we obtain

$$P(X > c_\alpha) = .983.$$

Once again we ask: How accurate are these approximations? Table 7 indicates that the approximations are quite good and later we shall see

Table 7: Monte Carlo Power Estimates for the Likelihood Ratio Test

		$H_o: \theta_1 = 0$ against $H_1: \theta_1 \neq 0$					$H_o: \theta_3 = -1$ against $H_1: \theta_3 \neq -1$				
Parameters*					Monte Carlo					Monte Carlo	
θ_1	θ_3	λ_1	λ_2	$P[X > c_\alpha]$	$P[L > F_\alpha]$	STD. ERR.	λ_1	λ_2	$P[X > c_\alpha]$	$P[L > F_\alpha]$	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	0.2423	0.0006	.103	.110	.004
0.015	-1.2	0.8307	0.0000	.237	.231	.006	0.8526	0.0078	.244	.248	.006
0.030	-1.4	3.3343	0.0000	.700	.687	.006	2.6928	0.0728	.622	.627	.007

* $\theta_2 = 1$, $\theta_4 = -.5$, $\sigma^2 = .001$

several more examples where this is the case. In general, Monte Carlo evidence suggests that the approximation $P(L > c_\alpha) \doteq P(X > c_\alpha)$ is very accurate over a wide range of circumstances. Table 7 was constructed exactly as Table 5.

In most applications λ_2 will be quite small relative to λ_1 as in the three cases in the last example. This being the case, one sees by scanning the entries in Table 6 that the value of $P(X > c_\alpha)$ computed with $\lambda_2 = 0$ would be adequate to approximate $P(L > F_\alpha)$. If $\lambda_2 = 0$ then (Problem 5)

$$H(c_\alpha;\ \nu_1,\ \nu_2,\ \lambda_1,\ 0) = F'(F_\alpha;\ \nu_1,\ \nu_2,\ \lambda_1)$$

with

$$c_\alpha = 1 + \nu_1 F_\alpha / \nu_2 .$$

Recall that $F'(x;\ \nu_1,\ \nu_2,\ \lambda)$ denotes the non-central F-distribution with ν_1 numerator degrees of freedom, ν_2 denominator degrees of freedom, and non-centrality parameter λ (Appendix 1). Stated differently, the first rows of Parts A through I of Table 6 are a tabulation of the power of the F-test. Thus, in most applications, an adequate approximation to the power of the likelihood ratio test is

$$P(L > F_\alpha) \doteq 1 - F'(F_\alpha;\ \nu_1,\ \nu_2,\ \lambda_1)$$

The next example explores the adequacy of this approximation.

Table 8. Monte-Carlo Power Estimates for an Exponential Model

Parameters		Non-centralities			Power Monte-Carlo	
θ_1	θ_2	λ_1	λ_2	$P[X > c_\alpha]$	$\hat{p}$	$SE(\hat{p})$
.5	.5	0	0	.050	.0532	.00308
.5398	.5	.9854	0	.204	.2058	.00570
.4237	.6849	.9853	.00034	.204	.2114	.00570
.5856	.5	4.556	0	.727	.7140	.00630
.3473	.8697	4.556	.00537	.728	.7312	.00629
.62	.5	8.958	0	.957	.9530	.00287

EXAMPLE 3. Table 8 compares the probability $P(X > c_\alpha)$ to Monte Carlo estimates of the probability of $P(L > F_\alpha)$ for the model

$$y_t = \theta_1 e^{\theta_2 x_t} + e_t.$$

Thirty inputs $\{x_t\}_{t=1}^{30}$ were chosen by replicating the points 0 (.1) .7 three times and the points .8 (.1) 1 twice. The null hypothesis is H: $\theta^o = (1/2, 1/2)$. For the null hypothesis and selected departures from the null hypothesis, 5000 random samples of size thirty from the normal distribution were generated according to the model with σ^2 taken as .04. The point estimate $\hat{p}$ of $P(L > F_\alpha)$ is, of course, the ratio of the number of times L exceeded F_α to 5000. The variance of $\hat{p}$ was estimated by $\mathrm{Var}(\hat{p}) = P(X > c_\alpha)\, P(X \leq c_\alpha)/5000$. For complete details see Gallant (1975a).

To comment on the choice of the values of $\theta^o \neq (1/2, 1/2)$ shown in Table 8, the ratio λ_2/λ_1 is minimized (=0) for $\theta^o \neq (1/2, 1/2)$ of the form $(\theta_1, 1/2)$ and is maximized for θ^o of the form $(1/2, 1/2) \pm r[\cos(5\pi/8), \sin(5\pi/8)]$. Three points were chosen to be of the first form and two of the latter form. Further, two sets of points were paired with respect to λ_1. This was done to evaluate the variation in power when λ_2 changes while λ_1 is held fixed.

These simulations indicate that the approximation of $P(L > F_\alpha)$ by $P(X > c_\alpha)$ is quite accurate as is the approximation

$$P(X > c_\alpha) \doteq 1 - F'(F_\alpha;\, q,\, n-p,\, \lambda_1).$$

EXAMPLE 2 (continued). As mentioned at the beginning of the chapter, the model

$$B\colon\ y_t = \theta_1 (e^{-\theta_2 x_t} - e^{-\theta_1 x_t})/(\theta_1 - \theta_2) + e_t$$

Table 9. Monte-Carlo Estimates of Power

		Wald Test			Likelihood Ratio		
		Monte-Carlo Estimate			Monte-Carlo Estimate		
$(\theta_1 - 1.4)/\sigma_1$	$(\theta_2 - .4)/\sigma_2$	$P[Y > F_\alpha]$	$P[W > F_\alpha]$	Std. Err.	$P[X > c_\alpha]$	$P[L > c_\alpha]$	Std. Err.
			a. Model B				
-4.5	1.0	.9725	.9835	.0017*	.9889	.9893	.0020
-3.0	0.5	.6991	.7158	.0027*	.7528	.7523	.0035
-1.5	-1.5	.2943	.2738	.0023*	.3051	.3048	.0017
1.5	-0.5	.2479	.2539	.0018*	.2379	.2379	.0016
3.0	-4.0	.9938	.9948	.0008	.9955	.9948	.0006
2.0	3.0	.7127	.7122	.0017*	.6829	.6800	.0028
-1.5	1.0	.3295	.3223	.0022*	.3381	.3368	.0015
0.5	-0.5	.0885	.0890	.0016	.0885	.0892	.0009
0.0	0.0	.0500	.0525	.0012*	.0500	.0501	.0008
			b. Model C				
-2.5	0.5	.9964	.9540	.0009*	1.0000	1.0000	.0000
-1.0	0.0	.5984	.4522	.0074*	.7738	.7737	.0060
2.0	-1.5	.4013	.4583	.0062*	.2807	.2782	.0071
0.5	-1.0	.2210	.2047	.0056*	.2877	.2892	.0041
4.5	-3.0	.9945	.8950	.0012*	.9736	.9752	.0025
0.0	1.0	.5984	.7127	.0054*	.5585	.5564	.0032
-2.0	3.5	.9795	.7645	.0022*	.4207	.4192	.0078*
-0.5	1.0	.2210	.3710	.0055*	.1641	.1560	.0040*
0.0	0.0	.0500	.1345	.0034*	.0500	.0502	.0012

Model B: $\sigma_1 = 0.052957$, $\sigma_2 = 0.014005$. Model C: $\sigma_1 = 0.27395$, $\sigma_2 = 0.029216$.

was chosen by Guttman and Meeter (1965) to represent a nearly linear model as measured by measures of the coincidence of the contours of $\|y - f(\theta)\|^2$ with the contours of $(\theta - \hat{\theta})'\hat{C}(\theta - \hat{\theta})$ introduced by Beale (1960). The model

$$C: y_t = 1 - (\theta_1 e^{-\theta_2 x_t} - \theta_2 e^{-\theta_1 x_t})/(\theta_1 - \theta_2) + e_t$$

is highly nonlinear by this same criterion. The simulations reported in Table 9 were designed to determine how the approximations

$$P(W > F_\alpha) \doteq P(Y > F_\alpha)$$

$$P(L > F_\alpha) \doteq P(X > c_\alpha)$$

hold up as we move from a nearly linear situation to more nonlinear situations. As we have hinted at all along, the approximation

$$P(W > F_\alpha) \doteq P(Y > F_\alpha)$$

deteriorates badly while the approximation

$$P(L > F_\alpha) \doteq P(X > c_\alpha)$$

holds up quite well. The details of the simulation are as follows.

The probabilities $P(W > F_\alpha)$ and $P(L > F_\alpha)$ that the hypothesis H: $\theta^o = (1.4, .4)$ is rejected shown in Table 9 were computed from 4000 Monte Carlo trials using the control variate method of variance reduction (Hammersly and Handscomb, 1964). The independent variables were the same as those listed

in Table 2 and the simulated errors were normally distributed with mean zero and variance $\sigma^2 = (.025)^2$. The sample size in each of the 4000 trials was $n = 12$ as one sees from Table 2. An asterisk indicates that $P(W > F_\alpha)$ is significantly different from $P(Y > F_\alpha)$ at the 5% level; similarly for the likelihood ratio test. For complete details see Gallant (1976).

If the null hypothesis is written as a parametric restriction

$$H: h(\theta^o) = 0$$

and it is not convenient to rewrite it as a functional dependency $\theta = g(\rho)$ the following alternative formula (Section 6 of Chapter 3) may be used to compute P_{FG}.

$$\theta_n^* \text{ minimizes } \sum_{t=1}^{n} [f(x_t,\theta^o) - f(x_t,\theta)]^2 \text{ subject to } h(\theta) = 0$$

$$\bar{H} = H(\theta_n^*) = (\partial/\partial\theta')h(\theta_n^*)$$

$$P_{FG} = P_F - F(F'F)^{-1}\bar{H}'[\bar{H}(F'F)^{-1}\bar{H}']^{-1}\bar{H}(F'F)^{-1}F'$$

We have discussed the Wald test and the likelihood test of

$$H: h(\theta^o) = 0 \quad \text{against} \quad A: h(\theta^o) \neq 0,$$

equivalently,

$$H: \theta^o = g(\rho) \text{ for some } \rho^o \quad \text{against} \quad A: \theta^o \neq g(\rho) \text{ for any } \rho$$

There is one other test in common use, the Lagrange multiplier (Problem 6) or efficient score test. In view of the foregoing, the following motivation is likely to have the strongest intuitive appeal. Let

$$\tilde{\theta} \text{ minimize } SSE(\theta) \text{ subject to } h(\theta) = 0,$$

equivalently,

$$\tilde{\theta} = g(\rho) \text{ where } \hat{\rho} \text{ minimizes } SSE[g(\rho)]$$

Suppose that $\tilde{\theta}$ is used as a starting value, the Gauss-Newton step away from $\tilde{\theta}$ (presumably) toward $\hat{\theta}$ is

$$\tilde{D} = (\tilde{F}'\tilde{F})^{-1}\tilde{F}'[y - f(\tilde{\theta})]$$

where $\tilde{F} = F(\tilde{\theta}) = (\partial/\partial\theta')f(\tilde{\theta})$. Intuitively, if the hypothesis $h(\theta^o) = 0$ is false then minimization of $SSE(\theta)$ subject to $h(\theta) = 0$ will cause a large displacement away from $\hat{\theta}$ and $\tilde{D}$ will be large. Conversely, if $h(\theta^o)$ is true then $\tilde{D}$ should be small. It remains to find some measure of the distance of $\tilde{D}$ from zero that will yield a convenient test statistic.

Recall that

$$\theta_n^* \text{ minimizes } \sum_{t=1}^{n} [f(x_t,\theta^o) - f(x_t,\theta)]^2 \text{ subject to } h(\theta) = 0,$$

equivalently,

$$\theta_n^* = g(\rho_n^o) \text{ where } \rho_n^o \text{ minimizes } \sum_{t=1}^{n} \{f(x_t,\theta^o) - f[x_t,g(\rho)]\}^2$$

and that

$$\delta = f(\theta^{o}) - f(\theta_{n}^{*})$$

$$P_{F} = F(F'F)^{-1}F'$$

$$P_{FG} = FG(G'F'FG)^{-1}G'F'$$

where $G = (\partial/\partial\rho')g(\rho_{n}^{o})$. Equivalently,

$$P_{FG} = P_{F} - F(F'F)^{-1}\bar{H}[\bar{H}(F'F)^{-1}\bar{H}']^{-1}\bar{H}(F'F)^{-1}F'$$

where $\bar{H} = (\partial/\partial\theta')h(\theta_{n}^{*})$. We shall show in Chapter 4 that

$$\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D}/n = (e + \delta)'(P_{F} - P_{FG})(e + \delta)/n + o_{p}(1/n),$$

$$SSE(\tilde{\theta})/n = (e + \delta)'(I - P_{FG})(e + \delta)/n + o_{p}(1/n),$$

$$SSE(\hat{\theta})/n = e'(I - P_{F})e/n + o_{p}(1/n).$$

These characterizations suggest two test statistics

$$R_{1} = \frac{\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D}/q}{SSE(\theta)/(n - p)}$$

and

$$R_2 = n \tilde{D}' (\tilde{F}' \tilde{F}) \tilde{D} / SSE(\tilde{\theta})$$

The second statistic R_2 is the customary form of the Lagrange multiplier test and has the advantage that it can be computed from knowledge of $\tilde{\theta}$ alone. The first requires two minimizations, one to compute $\hat{\theta}$ and another to compute $\tilde{\theta}$. Much is gained by going to this extra bother. The distribution theory is simpler and the test has better power as we shall see later on.

The two test statistics can be characterized as

$$R_1 = Z_1 + o_p(1)$$

$$R_2 = Z_2 + o_p(1)$$

where

$$Z_1 = \frac{(e + \delta)' (P_F - P_{FG})(e + \delta)/q}{e' (I - P_F)e/(n - p)}$$

$$Z_2 = n(e + \delta)' (P_F - P_{FG})(e + \delta)/(e + \delta)' (I - P_{FG})(e + \delta).$$

The distribution function of Z_1 is (Problem 7)

$$F'(z;\ q,\ n-p,\ \lambda_1)$$

where

$$\lambda_1 = \delta' (P_F - P_{FG})\delta/(2\sigma^2).$$

That is, the random variable Z_1 is distributed as the non-central F-distribution (Appendix 1) with q numerator degrees of freedom, n-p denominator degrees of freedom, and non-centrality parameter λ_1. Thus R_1 is approximately distributed as the (central) F distribution and the test is: Reject H when R_1 exceeds $F_\alpha = F^{-1}(1 - \alpha; q, n - p)$.

The distribution function of Z_2 is (Problem 8) for $z < n$

$$F''[(n-p)(z)/(q)(n-z); q, n-p, \lambda_1, \lambda_2]$$

where

$$\lambda_1 = \delta'(P_F - P_{FG})\delta/(2\sigma^2)$$

$$\lambda_2 = \delta'(I - P_F)\delta/(2\sigma^2).$$

and $F''(t; q, n-p, \lambda_1, \lambda_2)$ denotes the doubly non-central F-distribution (Appendix 1) with q numerator degrees of freedom, n-p denominator degrees of freedom, numerator non-centrality parameter λ_1 and denominator non-centrality parameter λ_2 (Appendix 1). If we approximate

$$P(R_2 > d) \doteq P(Z_2 > d)$$

then under the null hypothesis that $h(\theta^o) = 0$ we have $\delta = 0$, $\lambda_1 = 0$, and $\lambda_2 = 0$ whence

$$P(R_2 > d | \lambda_1 = \lambda_2 = 0) \doteq 1 - F[(n-p)(d)/(q)(n-d); q, n-p]$$

Letting F_α denote the $\alpha \times (100\%)$ critical point of the F-distribution, that is

$$\alpha = 1 - F(F_\alpha; q, n-p)$$

then that value d_α of d for which

$$P(R > d_\alpha | \lambda_1 = \lambda_2 = 0) = \alpha$$

is

$$F_\alpha = (n-p)(d_\alpha)/(q)(n-d_\alpha)$$

or

$$d_\alpha = nF_\alpha/[(n-p)/q + F_\alpha].$$

The test is then: Reject H: $h(\theta^o) = 0$ if $R_2 > d_\alpha$. With this computation of d_α,

$$P(R_1 > F_\alpha) \doteq P(Z_1 > F_\alpha)$$

$$= 1 - F'(F_\alpha; q, n-p, \lambda_1)$$

$$\leq 1 - F''(F_\alpha; q, n-p, \lambda_1, \lambda_2)$$

$$= P(Z_2 > d_\alpha)$$

$$\doteq P(R_2 > d_\alpha)$$

and we see that to within the accuracy of these approximations, the first version of the Lagrange multiplier test always has better power than the second. Of course as we noted earlier, in most instances λ_2 will be small relative to λ_1 and the difference in power will be negligible.

In the same vein, judging from the entries in Table 6 we have (see Problem 10)

$$1 - F'(F_\alpha; q, n-p, \lambda_1) \leq 1 - H(c_\alpha; q, n-p, \lambda_1, \lambda_2)$$

whence

$$P(L > F_\alpha) \doteq P(X > c_\alpha)$$

$$= 1 - H(c_\alpha; q, n-p, \lambda_1, \lambda_2)$$

$$\geq 1 - F'(F_\alpha; q, n-p, \lambda_1)$$

$$= P(Z_1 > F_\alpha)$$

$$\doteq P(R_1 > F_\alpha).$$

Thus the likelihood ratio test has better power than either of the two versions of the Lagrange multiplier test. But again, λ_2 is usually small and the difference in power negligible.

To summarize this discussion, the first version of the Lagrange multiplier test rejects the hypothesis

$$H: h(\theta^{o}) = 0$$

when the statistic

$$R_1 = \frac{\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D}/q}{SSE(\tilde{\theta})/(n-p)}$$

exceeds $F_\alpha = F^{-1}(1-\alpha;\ q,\ n-p)$. The second version rejects when the statistic

$$R_2 = n\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D}/SSE(\tilde{\theta})$$

exceeds

$$d_\alpha = nF_\alpha/[(n-p)/q + F_\alpha].$$

As usual, there are various strategies one might employ to compute the statistics R_1 and R_2. In connection with the likelihood ratio test, we have already discussed and illustrated how one can compute $\tilde{\theta}$ by computing the unconstrained minimum $\hat{\rho}$ of the composite function $SSE[g(\rho)]$ and setting $\tilde{\theta} = g(\hat{\rho})$. Now suppose that one creates a data set with observations

$$\tilde{e}_t = y_t - f(x_t, \tilde{\theta}) \qquad t=1,2,\ldots,n$$

$$\tilde{f}_t' = (\partial/\partial\theta')f(x_t, \tilde{\theta}) \qquad t=1,2,\ldots,n$$

Or in vector notation

$$\tilde{e} = y - f(\tilde{\theta}), \quad \tilde{F} = (\partial/\partial\theta')f(\tilde{\theta})$$

Note that $\tilde{F}$ is an n by p matrix; $\tilde{F}$ is <u>not</u> the n by r matrix $(\partial/\partial\rho')f[g(\tilde{\rho})]$. If one regresses $\tilde{e}$ on $\tilde{F}$ with no intercept term using a linear regression procedure then the analyssis of variance table printed by the program will have the following entries

Source	d.f.	Sum of Squares
Regression	p	$\tilde{e}'\tilde{F}(\tilde{F}'\tilde{F})^{-1}\tilde{F}'\tilde{e}$
Error	n-p	$\tilde{e}'\tilde{e} - \tilde{e}'\tilde{F}(\tilde{F}'\tilde{F})\tilde{F}'\tilde{e}$
Total	n	$\tilde{e}'\tilde{e}$

One can just read off

$$\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D} = \tilde{e}'\tilde{F}(\tilde{F}'\tilde{F})^{-1}\tilde{F}'\tilde{e}$$

$$SSE(\tilde{\theta}) = \tilde{e}'\tilde{e}$$

from the analysis of variance table. Let us illustrate these ideas.

EXAMPLE 1 (continued). Recalling that the response function is

$$f(x,\theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}$$

reconsider the first hypothesis

Fisure 12a. Illustration of Lasranse Multiplier Test Computations with Example 1.

SAS Statements:

```
DATA WORK01;  SET EXAMPLE1;
T1=0.0;  T2=1.00296592;  T3=-1.14123442;  T4=-0.51182277;
E=Y-(T1*X1+T2*X2+T4*EXP(T3*X3));
F1=X1;  F2=X2;  F3=T4*X3*EXP(T3*X3);  F4=EXP(T3*X3);
DROP T1 T2 T3 T4;
PROC REG DATA=WORK01;  MODEL E=F1 F2 F3 F4 / NOINT;
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

DEP VARIABLE: E

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	4	0.004938382	0.001234596	1.053	0.3996
ERROR	26	0.030495	0.001172869		
U TOTAL	30	0.035433			

ROOT MSE	0.034247	R-SQUARE	0.1394
DEP MEAN	-5.50727E-09	ADJ R-SQ	0.0401
C.V.	-621854289		

NOTE: NO INTERCEPT TERM IS USED. R-SQUARE IS REDEFINED.

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
F1	1	-0.025888	0.012616	-2.052	0.0504
F2	1	0.012719	0.009874181	1.288	0.2091
F3	1	0.026417	0.165440	0.160	0.8744
F4	1	0.007033215	0.025929	0.271	0.7883

$$H: \theta_1^o = 0.$$

Previously we computed

$$\tilde{\theta} = \begin{pmatrix} 0.0 \\ 1.00296592 \\ -1.14123442 \\ -0.51182277 \end{pmatrix} \qquad \text{(from Figure 9a)}$$

$$SSE(\tilde{\theta}) = 0.03543298 \qquad \text{(from Figure 9a or Figure 12a)}$$

$$SSE(\hat{\theta}) = 0.03049554 \qquad \text{(from Figure 5a)}$$

We implement the scheme of regressing $\tilde{e}$ on $\tilde{F}$ in Figure 12a (note the similarity with Figure 11a) and obtain

$$\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D} = 0.004938382 \qquad \text{(from Figure 12a)}$$

The first Lagrange multiplier test statistic is

$$R_1 = \frac{\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D}/q}{SSE(\tilde{\theta})/(n-p)}$$

$$= \frac{(0.004938382)/(1)}{(0.03049554)/(26)}$$

$$= 4.210.$$

Comparing with the critical point

$$F^{-1}(.95;\ 1,\ 26) = 4.22$$

one fails to reject the null hypothesis at the 95% level.

The second Lagrange multiplier test statistic is

$$R_2 = n\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D}/SSE(\tilde{\theta})$$

$$= (30)(0.004938382)/(0.03543298)$$

$$= 4.1812$$

Comparing with the critical point

$$d_\alpha = nF_\alpha/[(n-p)/q + F_\alpha]$$

$$= (30)(4.22)/[(26)/(1) + 4.22]$$

$$= 4.19$$

One fails to reject the null hypothesis at the 95% level.

Reconsider the second hypothesis

$$H:\ \theta_3\theta_4 e^{\theta_3} = 1/5$$

which can be represented equivalently as

Figure 12b. Illustration of Lagrange Multiplier Test Computations with Example 1.

SAS Statements:

```
DATA WORK01;  SET EXAMPLE1;
R1=-0.02301828;  R2=1.01965423;  R3=-1.16048699;
T1=R1; T2=R2; T3=R3; T4=1/(5*R3*EXP(R3));
E=Y-(T1*X1+T2*X2+T4*EXP(T3*X3));
F1=X1;  F2=X2;  F3=T4*X3*EXP(T3*X3);  F4=EXP(T3*X3);
DROP T1 T2 T3 T4;
PROC REG DATA=WORK01;  MODEL E=F1 F2 F3 F4 / NOINT;
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

DEP VARIABLE: E

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	4	0.004439308	0.001109827	0.946	0.4531
ERROR	26	0.030493	0.001172804		
U TOTAL	30	0.034932			

ROOT MSE	0.034246	R-SQUARE	0.1271
DEP MEAN	7.59999E-09	ADJ R-SQ	0.0264
C.V.	450609078		

NOTE: NO INTERCEPT TERM IS USED. R-SQUARE IS REDEFINED.

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
F1	1	-0.00285742	0.012611	-0.227	0.8225
F2	1	-0.00398546	0.009829362	-0.405	0.6885
F3	1	0.043503	0.156802	0.277	0.7836
F4	1	0.045362	0.026129	1.736	0.0944

$$H: \theta^{o} = g(\rho) \text{ for some } \rho^{o}$$

with

$$g(\rho) = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ 1/(5\rho_3 e^{\rho_3}) \end{pmatrix}$$

Previously we computed

$$\tilde{\rho} = \begin{pmatrix} -0.02301828 \\ 1.01965423 \\ -1.16048699 \end{pmatrix} \qquad \text{(from Figure 9b)}$$

$$SSE(\tilde{\theta}) = 0.03493222 \qquad \text{(from Figure 9b or Figure 12b)}$$

$$SSE(\hat{\theta}) = 0.03049554 \qquad \text{(from Figure 5a)}$$

Regressing $\tilde{e}$ on $\tilde{F}$ we obtain

$$\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D} = 0.004439308 \qquad \text{(from Figure 12b)}$$

The first Lagrange multiplier test statistic is

$$R_1 = \frac{\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D}/q}{SSE(\hat{\theta})/(n-p)}$$

$$= \frac{(0.004439308)/(1)}{(0.03049554)/(26)}$$

$$= 3.7849$$

Comparing with

$$F(.95;\ 1,\ 26) = 4.22$$

we fail to reject the null hypothesis at the 95% level.

The second Lagrange multiplier test statistic is

$$R_2 = n\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D}/SSE(\tilde{\theta})$$

$$= (30)(0.004439308)/(0.0349322)$$

$$= 3.8125$$

Comparing with

$$d_\alpha = nF_\alpha/[(n-p)/q + F_\alpha]$$

$$= (30)(4.22)/[(26)/(1) + 4.22]$$

$$= 4.19$$

we fail to reject at the 95% level.

Reconsidering the third hypothesis

$$H: \theta_1 = 0 \quad \text{and} \quad \theta_3\theta_4 e^{\theta_3} = 1/5$$

which may be rewritten as

$$H: \theta^o = g(\rho) \quad \text{for some} \quad \rho^o$$

with

$$g(\rho) = \begin{pmatrix} 0 \\ \rho_2 \\ \rho_3 \\ 1/(5\rho_3 e^{\rho_3}) \end{pmatrix}$$

Previously we computed

$$\tilde{\theta} = \begin{pmatrix} \tilde{\rho}_2 \\ \tilde{\rho}_3 \end{pmatrix} = \begin{pmatrix} 1.00795735 \\ -1.16918683 \end{pmatrix} \qquad \text{(from Figure 9c)}$$

$$SSE(\tilde{\theta}) = 0.03889923 \qquad \text{(from Figure 9c or Figure 12c)}$$

$$SSE(\tilde{\theta}) = 0.03049554 \qquad \text{(from Figure 5a)}$$

Regressing $\tilde{e}$ on $\tilde{F}$ we obtain

$$\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D} = 0.008407271 \qquad \text{(from Figure 12c)}$$

Figure 12c. Illustration of Lagrange Multiplier Test Computations with Example 1.

SAS Statements:

```
DATA WORK01;  SET EXAMPLE1;
R1=0;  R2=1.00795735;  R3=-1.16918683;
T1=R1; T2=R2; T3=R3; T4=1/(5*R3*EXP(R3));
E=Y-(T1*X1+T2*X2+T4*EXP(T3*X3));
F1=X1;  F2=X2;  F3=T4*X3*EXP(T3*X3);  F4=EXP(T3*X3);
DROP T1 T2 T3 T4;
PROC REG DATA=WORK01;  MODEL E=F1 F2 F3 F4 / NOINT;
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

DEP VARIABLE: E

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	4	0.008407271	0.002101818	1.792	0.1607
ERROR	26	0.030492	0.001172768		
U TOTAL	30	0.038899			

ROOT MSE	0.034246	R-SQUARE	0.2161
DEP MEAN	-2.83174E-09	ADJ R-SQ	0.1257
C.V.	-1209350370		

NOTE: NO INTERCEPT TERM IS USED. R-SQUARE IS REDEFINED.

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
F1	1	-0.025868	0.012608	-2.052	0.0504
F2	1	0.007699193	0.00980999	0.785	0.4396
F3	1	0.052092	0.157889	0.330	0.7441
F4	1	0.046107	0.026218	1.759	0.0904

The first Lagrange multiplier test statistic is

$$R_1 = \frac{\tilde{D}'(\tilde{F}'\tilde{F})D/q}{SSE(\theta)/(n-p)}$$

$$= \frac{(0.008407271)/(2)}{(0.03049554)/(26)}$$

$$= 3.5840$$

Comparing with

$$F^{-1}(.95;\ 2,\ 26) = 3.37$$

we reject the null hypothesis at the 5% level.

The second Lagrange multiplier test statistic is

$$R_2 = n\tilde{D}'(\tilde{F}'\tilde{F})D/SSE(\tilde{\theta})$$

$$= (30)(0.008407271)/(0.03889923)$$

$$= 6.4839$$

Comparing with

$$d_\alpha = nF_\alpha[(n - p)/q + F_\alpha]$$

$$= (30)(3.37)/[(26)/2 + 3.37]$$

$$= 6.1759$$

we reject at the 95% level. ∎

As the example suggests, the approximation

$$\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D} \doteq SSE(\tilde{\theta}) - SSE(\hat{\theta})$$

is quite good so that

$$R_1 \doteq L$$

in most applications. Thus, in most instances, the likelihood ratio test and the first version of the Lagrange multiplier test will accept and reject together.

To compute power, one uses the approximations

$$P(R_1 > F_\alpha) \doteq P(Z_1 > F_\alpha)$$

and

$$P(R_2 > d_\alpha) \doteq P(Z_2 > d_\alpha).$$

The non-centrality parameters λ_1, and λ_2 appearing in the distributions of Z_1 and Z_2 are the same as those in the distribution of X. Their computation was discussed in detail during the discussion of power computations for the likelihood ratio test. We illustrate

EXAMPLE 1 (continued). Recalling that

$$f(x,\theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}$$

let us approximate the probabilities that the two versions of the Lagrange multiplier test reject the following three hypotheses at the 5% level when the true values of the parameters are

$$\theta^{o} = (.03,\ 1,\ -1.4,\ -.5)'$$

$$\sigma^2 = .001$$

The three hypotheses are the same as those we have used for the illustration throughout:

$$H_1\colon\ \theta_1 = 0$$

$$H_2\colon\ \theta_3\theta_4 e^{\theta_3} = 1/5$$

$$H_3\colon\ \theta_3 = 0 \quad \text{and} \quad \theta_3\theta_4 e^{\theta_3} = 1/5.$$

In connection with the illustration of power computations for the likelihood ratio test we obtained

$$H_1\colon\ \lambda_1 = 3.3343, \qquad \lambda_2 = 0$$

$$H_2\colon\ \lambda_1 = 6.5128 \qquad \lambda_2 = 0.0005827$$

$$H_3\colon\ \lambda_1 = 10.9604 \qquad \lambda_2 = 0.0008241.$$

For the first hypothesis

$$P(R_1 > F_\alpha) \doteq P(Z_1 > F_\alpha)$$

$$= 1 - F'(F_\alpha;\ q,\ n-p,\ \lambda_1)$$

$$= 1 - F'(4.22;\ 1,\ 26,\ 3.3343)$$

$$= .700\ ,$$

$$P(R_2 > d_\alpha) \doteq P(Z_2 > d_\alpha)$$

$$= 1 - F''(F_\alpha;\ q,\ n-p,\ \lambda_1,\ \lambda_2)$$

$$= 1 - F''(4.22;\ 1,\ 26,\ 3.3343,\ 0)$$

$$= .700\ ;$$

the second

$$P(R_1 > F_\alpha) \doteq P(Z_1 > F_\alpha)$$

$$= 1 - F'(F_\alpha;\ q,\ n-p,\ \lambda_1)$$

$$= 1 - F'(4.22;\ 1,\ 26,\ 6.5128)$$

$$= .935$$

$$P(R_2 > d_\alpha) \doteq P(Z_2 > d_\alpha)$$

$$= 1 - F''(F_\alpha;\ q,\ n-p,\ \lambda_1,\ \lambda_2)$$

$$= 1 - F''(4.22;\ 1,\ 26,\ 6.5128,\ 0.0005827)$$

$$= .935$$

and the third

$$P(R_1 > F_\alpha) \doteq P(Z_1 > F_\alpha)$$

$$= 1 - F'(F_\alpha;\ q,\ n-p,\ \lambda_1)$$

$$= 1 - F'(3.37;\ 2,\ 26,\ 10.9604)$$

$$= .983$$

$$P(R_2 > d_\alpha) \doteq P(Z_2 > d_\alpha)$$

$$= 1 - F''(F_\alpha;\ q,\ n-p,\ \lambda_1,\ \lambda_2)$$

$$= 1 - F''(3.37;\ 2,\ 26,\ 10.9604,\ 0.0008241)$$

$$= .983$$

Table 10a. Monte Carlo Power Estimates for Version 1 of the Lagrange Multiplier Test

H_o: $\theta_1 = 0$ against H_1: $\theta_1 \neq 0$							H_o: $\theta_3 = -1$ against H_1: $\theta_3 \neq -1$				
Parameters*					Monte Carlo					Monte Carlo	
θ_1	θ_3	λ_1	λ_2	$P[Z_1 > F_\alpha]$	$P[R_1 > F_\alpha]$	STD. ERR.	λ_1	λ_2	$P[Z_1 > F_\alpha]$	$P[R_1 > F_\alpha]$	STD. ERR.
0.0	-1.0	0.0	0.0	.050	.049	.003	0.0	0.0	.050	.051	.003
0.008	-1.1	0.2353	0.0000	.101	.094	.004	0.2423	0.0006	.103	.107	.004
0.015	-1.2	0.8307	0.0000	.237	.231	.006	0.8526	0.0078	.242	.241	.006
0.030	-1.4	3.3343	0.0000	.700	.687	.006	2.6928	0.0728	.608	.608	.007

* $\theta_2 = 1$, $\theta_4 = -.5$, $\sigma^2 = .001$

1-5-76

Again one questions the accuracy of these approximations. Tables 10a and 10b indicate that the approximations are quite good. Also, by comparing Tables 7, 10a and 10b one can see the beginnings of the spread

$$P(L > F_\alpha) > P(R_1 > F_\alpha) > P(R_2 > d_\alpha)$$

as λ_2 increases which was predicted by the theory. Tables 9a and 9b were constructed exactly the same as Tables 5 and 7.

Table 10b. Monte Carlo Power Estimates for Version 2 of the Lagrange Multiplier Test

		H_o: $\theta_1 = 0$ against H_1: $\theta_1 \neq 0$					H_o: $\theta_3 = -1$ against H_1: $\theta_3 \neq -1$				
Parameters*					Monte Carlo					Monte Carlo	
θ_1	θ_3	λ_1	λ_2	$P[Z_2 > d_\alpha]$	$P[R_2 > d_\alpha]$	STD. ERR.	λ_1	λ_2	$P[Z_2 > d_\alpha]$	$P[R_2 > d_\alpha]$	STD. ERR.
0.0	-1.0	0.0	0.0	.050	.049	.003	0.0	0.0	.050	.050	.003
0.008	-1.1	0.2353	0.0000	.101	.094	.004	0.2423	0.0006	.103	.106	.004
0.015	-1.2	0.8307	0.0000	.237	.231	.006	0.8526	0.0078	.242	.241	.006
0.030	-1.4	3.3343	0.0000	.700	.687	.006	2.6928	0.0728	.606	.605	.007

* $\theta_2 = 1$, $\theta_4 = -.5$, $\sigma^2 = .001$

1-5-78

PROBLEMS

1. Assuming that the density of y is $p(y;\ \theta,\sigma) = (2\pi\sigma^2)^{-n/2} \exp\{-(1/2)[y - f(\theta)]'[y - f(\theta)]/\sigma^2\}$ show that

$$\max_{\theta,\sigma} p(y;\ \theta,\ \sigma) = [2\pi SSE(\hat{\theta})/n]^{-n/2}\exp(-n/2)$$

$$\max_{h(\theta)=0,\sigma} p(y;\ \theta,\ \sigma) = [2\pi SSE(\tilde{\theta})/n]^{-n/2}\exp(-n/2),$$

presuming, of course, that $f(\theta)$ is such that the maximum exists. The likelihood ratio test rejects when the ratio

$$[\max_{h(\theta)=0,\sigma} p(y;\ \theta,\ \sigma)]/[\max_{\theta,\sigma} p(y;\ \theta,\ \sigma)]$$

is small. Put this statistic in the form: Reject when

$$\frac{[SSE(\tilde{\theta}) - SSE(\hat{\theta})]/q}{SSE(\hat{\theta})/(n-p)}$$

is large.

2. If the system of equations defined over Θ

$$h(\theta) = \tau$$

$$\phi(\theta) = \rho$$

has an inverse

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

show that

$$\{\theta \in \Theta: h(\theta) = 0\}$$

$$= \{\theta: \theta = \psi(\rho, 0) \text{ for some } \rho \text{ in } R\}$$

where $R = \{\rho: \rho = \phi(\theta) \text{ for some } \theta \text{ in } \Theta\}$.

3. Referring to the previous problem, show that

$$\max\{SSE(\theta): h(\theta) = 0 \text{ and } \theta \text{ in } \Theta\}$$

$$= \max\{SSE[\psi(\rho, 0)]: \rho \text{ in } R\}$$

if either maximum exists.

4. (Derivation of $H(x; \nu_1, \nu_2, \lambda_1, \lambda_2)$). Define $H(x; \nu_1, \nu_2, \lambda_1, \lambda_2)$ to be the distribution function given by

$$
\begin{cases}
0, & x \leq 1,\ \lambda_2 = 0, \\
\int_0^\infty G[t/(x-1) + 2x\lambda_2/(x-1)^2;\ \nu_2,\ \lambda_2/(x-1)^2] & \\
\qquad \cdot g(t;\ \nu_1,\ \lambda_1)dt, & x < 1,\ \lambda_2 > 0, \\
\int_0^\infty N(-t;\ 2\lambda_2,\ 8\lambda_2)g(t;\ \nu_1,\ \lambda_1)dt, & x = 1,\ \lambda_2 > 0, \\
1 - \int_0^\infty G[t/(x-1) + 2x\lambda_2/(x-1)^2;\ \nu_2,\ \lambda_2/(x-1)^2] & \\
\qquad \cdot g(t;\ \nu_1,\ \lambda_1)dt, & x > 1.
\end{cases}
$$

where $g(t;\ \nu,\ \lambda)$ denotes the non-central chi-square density function with ν degrees of freedom and non-centrality parameter λ and $G(t;\ \nu,\ \lambda)$ denotes the corresponding distribution function (Appendix 1).

Fill in the missing steps. Set $z = (1/\sigma)e$, $\gamma = (1/\sigma)\delta_0$, 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 $(z + b\gamma)'R(Z + b\gamma)$ is a noncentral chi-squared with q degrees freedom and noncentrality $b^2\gamma'R\gamma/2$, since R is idempotent with rank q. Similarly, $(z + b\gamma)'P^{\perp}(z + b\gamma)$ is a noncentral chi-squared with n - p degrees freedom and noncentrality $b^2\gamma'P^{\perp}\gamma/2$. These two random variables are independent because $RP^{\perp} = 0$.

Let $a > 0$.

$$
\begin{aligned}
P[X > a + 1] &= P[(z + \gamma)'P_1^{\perp}(z + \gamma) > (a + 1)z'P^{\perp}z] \\
&= P(z + \gamma)'R(z + \gamma) > az'P^{\perp}z - 2\gamma'P^{\perp}z - \gamma'P^{\perp}\gamma] \\
&= P[(z + \gamma)'R(z + \gamma) > a(z - a^{-1}\gamma)'P^{\perp}(z - a^{-1}\gamma) - (1 + a^{-1})\gamma'P^{\perp}\gamma]
\end{aligned}
$$

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

$$- (1 + a^{-1})\gamma'P^{\perp}\gamma]g(t;\ q,\ \gamma'R\gamma/2)dt$$

$$= \int_0^\infty P[(z - a^{-1}\gamma)'P^{\perp}(z - a^{-1}\gamma)$$

$$< (t + (1 + a^{-1})\gamma'P^{\perp}\gamma)/a]g(t;\ q,\ \gamma'R\gamma/2)dt$$

$$= \int_0^\infty G[t/a + (a + 1)\gamma'P^{\perp}\gamma/a^2;\ n - p,\ \gamma'P^{\perp}y/(2a^2)]g(t;\ q,\ \gamma'R\gamma/2)dt.$$

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

The derivations for the remaining cases are analogous.

5. Show that if $\lambda_2 = 0$, then

$$P(X > c_\alpha) = P[(n-p)(e + \delta)'(P_F - P_{FG})(e + \delta)/(qe'P_F^{\perp}e) > F_\alpha].$$

Referring to Problem 4, why doees this fact imply that

$$H(c_\alpha;\ \nu_1,\ \nu_2,\ \lambda_1,\ 0) = F'(F_\alpha;\ \nu_1,\ \nu_2,\ \lambda_1)\ ?$$

6. (Alternative motivation of the Lagrange multiplier test). Suppose that we change the sign conventions on the components of the vector valued function $h(\theta)$ so that

minimize $SSE(\theta)$

subject to $h(\theta) \leq 0$

is equivalent to the problem

minimize $SSE(\theta)$

subject to $h(\theta) = 0$.

The vector inequality means inequality component by component.

Now consider the problem

minimize $SSE(\theta)$

subject to $h(\theta) = x$

and view the solution θ as depending on x. Under suitable regularity conditions there is a vector x of Lagrange multipliers such that

$$(\partial/\partial\theta')SSE(\tilde{\theta}) = \tilde{\lambda}'H(\tilde{\theta})$$

and $(\partial/\partial x')\tilde{\theta}(x)$ exists. Then

$$h[\tilde{\theta}(x)] = x$$

implies

$$H(\tilde{\theta})(\partial/\partial x')\tilde{\theta}(x) = I$$

whence

$$(\partial/\partial x')SSE[\tilde{\theta}(x)]$$

$$= (\partial/\partial\theta')SSE[\tilde{\theta}(x)](\partial/\partial x')\tilde{\theta}(x)$$

$$= \tilde{\lambda}'H[\tilde{\theta}(x)](\partial/\partial x')\tilde{\theta}(x)$$

$$= \tilde{\lambda}'.$$

The intuitive interpretation of this equation is that if one had one more unit of the constraint h_i then $SSE(\theta)$ would increase by the amount λ_i. Then one should be willing to pay λ_i (in units of SSE) for one more unit of h_i. Stated differently, the components of the vector λ can be viewed as the prices of the constraints.

With this interpretation any reasonable measure $d(\lambda)$ of the distance of the vector $\tilde{\lambda}$ from zero could be used to test

$$H: h(\theta) = 0 \quad \text{against} \quad A: h(\theta) \neq 0.$$

One would reject for large values of $d(\tilde{\lambda})$. Show that if

$$d(\tilde{\lambda}) = (1/4)\tilde{\lambda}'\tilde{H}(\tilde{F}'\tilde{F})^{-1}\tilde{H}'\tilde{\lambda}$$

is chosen as the measure of distance where $\tilde{H}$ and $\tilde{F}$ denote evaluation of $\theta = \tilde{\theta}$ then

$$d(\tilde{\lambda}) = \tilde{D}'(\tilde{F}'\tilde{F})\tilde{D}$$

where, recall, $\tilde{D} = (\tilde{F}'\tilde{F})^{-1}\tilde{F}'[y - f(\tilde{\theta})]$.

7. Show that Z_1 is distributed as $F'(z;\ q,\ n\text{-}p,\ \lambda_1)$. Hint: $P_F(I - P_F) = 0$ and $P_{FG}(I - P_F) = 0$.

8. Fill in the missing steps. If $z < n$

$$P(Z_2 < z)$$

$$= P[(e + \delta)'(P_F - P_{FG})(e + \delta) < (z/n)(e + \delta)'(I - P_{FG})(e + \delta)]$$

$$= P\left[\frac{(e + \delta)'(P_F - P_{FG})(e + \delta)/q}{(e + \delta)'P_F^{\perp}(e + \delta)/(n - p)} < \frac{(n - p)z}{q(n - z)}\right]$$

$$= F''[(n - p)(z)/(q)(n - z);\ q,\ n\text{-}p,\ \lambda_1,\ \lambda_2].$$

9. (Relaxation of the Normality Assumption). The distribution of e is spherical if the distribution of Qe is the same as the distribution of e for every n by n orthogonal matrix Q. Perhaps the most useful distribution of this sort other than the normal is the multivariate Student-t (Zellner, 1976). Show that the null distributions of X, Z_1, and Z_2 do not change if any spherical distribution is substituted for the normal distribution. Hint: Jensen (1981).

10. Prove that $P(X > c_\alpha) \geq P(Z_1 > F_\alpha)$. Warning: this is an open question!

6. CONFIDENCE INTERVALS

A confidence interval on any (twice continuously differentiable) parametric function $\gamma(\theta)$ can be obtained by inverting any of the tests of

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

described in the previous section. That is, to construct a $100 \times (1-\alpha)\%$ confidence interval for $\gamma(\theta)$ one lets

$$h(\theta) = \gamma(\theta) - \gamma^{o}$$

and puts in the interval all those γ^{o} for which the hypothesis $H: h(\theta) = 0$ is accepted at the α level of significance (Problem 1). The same is true for confidence regions, the only difference being that $\gamma(\theta)$ and γ^{o} will be q-vectors instead of being univariate.

The Wald test is easy to invert. In the univariate case (q=1), the Wald test accepts when

$$|\gamma(\hat{\theta}) - \gamma^{o}|/(s^{2}\hat{H}\hat{C}\hat{H}')^{1/2} \leq t_{\alpha/2}$$

where

$$\hat{H} = (\partial/\partial\theta')[\gamma(\hat{\theta}) - \gamma^{o}] = (\partial/\partial\theta')\gamma(\hat{\theta})$$

and $t_{\alpha/2} = t^{-1}(1 - \alpha/2; n-p)$; that is, $t_{\alpha/2}$ denotes the upper $\alpha/2$ critical

point of the t-distribution with n-p degrees of freedom. Those points γ^o that satisfy the inequality are in the interval

$$\gamma(\hat{\theta}) \pm t_{\alpha/2}(s^2\hat{H}\hat{C}\hat{H}')^{1/2}.$$

The most common situation is when one wishes to set a confidence interval on one of the components θ_i of the parameter vector θ. In this case the interval is

$$\theta_i \pm t_{\alpha/2}\sqrt{s^2\hat{c}_{ii}}$$

where $\hat{c}_{ii}$ is the i-th diagonal element of $\hat{C} = [F'(\hat{\theta})F(\hat{\theta})]^{-1}$. We illustrate with Example 1.

EXAMPLE 1 (continued). Recalling that

$$f(x,\theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}$$

let us set a confidence interval on θ_1 by inverting the Wald test. One can read off the confidence interval directly from the SAS output of Figure 5a as

[-0.05183816, 0.00005877]

or compute it as

$\theta_1 = -0.02588970$ (from Figure 5a)

$\hat{c}_{11} = .13587$ (from Figure 5b)

$s^2 = 0.00117291$ (from Figure 5b)

$t^{-1}(.975;\ 26) = 2.0555$

$$\hat{\theta}_1 \pm t_{\alpha/2}\sqrt{s^2\hat{c}_{11}} = -0.02588970 \pm (2.0555)\sqrt{(0.00117291)(.13587)}$$
$$= -0.02588970 \pm 0.0259484615$$

whence

$$[-0.051838,\ 0.000588].$$

To put a confidence interval on

$$\gamma(\theta) = (\partial/\partial x_3)f(x,\theta)\Big|_{x_3=1} = \theta_3\theta_4 e^{\theta_3}$$

we have

$$H(\theta) = (\partial/\partial\theta')\gamma(\theta) = [0,\ 0,\ \theta_4(1+\theta_3)e^{\theta_3},\ \theta_3 e^{\theta_3}]$$

$\gamma(\theta) = (-1.11569714)(-0.50490286)e^{-1.11569714}$ (from Figure 5a)

$= 0.1845920697$

$H = (0,\ 0,\ 0.0191420895),\ -0.365599176)$ (from Figure 5a)

$\hat{H}\hat{C}\hat{H}' = 0.0552562$ (from Figures 5b and 13)

Figure 13. Wald Test Confidence Interval Construction Illustrated with Example 1.

SAS Statements:

```
PROC MATRIX;
C =  0.13587   -0.067112    -0.15100   -0.037594/
   -0.067112    0.084203     0.51754 -0.00157848/
    -0.15100     0.51754     22.8032     2.00887/
   -0.037594 -0.00157848     2.00887     0.56125;
H =  0  0  0.0191420895  -0.365599176;
HCH = H*C*H' ;  PRINT HCH;
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

	HCH	COL1
ROW1		0.0552563

$$s^2 = 0.00117291 \qquad \text{(from Figure 5a)}$$

Then the confidence interval is

$$\gamma(\hat{\theta}) \pm t_{\alpha/2}(s^2 \hat{H}\hat{C}\hat{H})^{1/2}$$

$$= 0.184592 \pm (2.0555)[(0.00117291)(0.0552563)]^{1/2}$$

$$= 0.1845921 \pm 0.0165478$$

or

$$[0.168044,\ 0.201140]. \quad \blacksquare$$

In the case that $\gamma(\theta)$ is a q-vector, the Wald test accepts when

$$[\gamma(\hat{\theta}) - \gamma^o]'(HCH')^{-1}[\gamma(\hat{\theta}) - \gamma^o]/(qs^2) \leq F_\alpha.$$

The confidence region obtained by inverting this test is an ellipsoid with center at $\gamma(\theta)$ and the eigenvectors of $\hat{H}\hat{C}\hat{H}'$ as axes.

To construct a confidence interval for $\gamma(\theta)$ by inverting the likelihood ratio test, put

$$h(\theta) = \gamma(\theta) - \gamma^o$$

with γ^o being a q-vector and let

$$SSE_{\gamma^o} = \min\{SSE(\theta): \gamma(\theta) = \gamma^o\}$$

The likelihood ratio test accepts when

$$L(\gamma^o) = \frac{(SSE_{\gamma^o} - SSE_{full})/q}{(SSE_{full})/(n-p)} \leq F_\alpha$$

where, recall, $F_\alpha = F^{-1}(1-\alpha;\ q,\ n-p)$ and $SSE_{full} = SSE(\hat{\theta}) = \min SSE(\theta)$. Thus, a likelihood ratio confidence region consists of those points γ^o with $L(\gamma^o) \leq F_\alpha$. Although it is not a frequent occurrence in applications, the likelihood ratio test can have unusual structural characteristics. It is possible that $L(\gamma^o)$ does not rise above F_α as $\|\gamma^o\|$ increases in some diretion so that the confidence region can be unbounded. Also it is possible that $L(\gamma^o)$ has local minima which can lead to confidence regions consisting of disjoint islands. But as we said, this does not happen often.

In the univariate case, the easiest way to invert the likelihood ratio test is by quadratic interpolation as follows. Take three trial values γ_1^o, γ_2^o, γ_3^o around the lower limit of the Wald test confidence interval and compute the corresponding values of $L(\gamma_1^o)$, $L(\gamma_2^o)$, $L(\gamma_3^o)$. Fit the quadratic equation

$$L(\gamma_i^o) = a(\gamma_i^o)^2 + b(\gamma_i^o) + c \qquad i=1,2,3$$

to these three points and let $\hat{x}$ solve the equation

$$F_\alpha = ax^2 + bx + c$$

One can take $\hat{x}$ as the lower limit or refine the estimates by taking three

trial values γ_1^o, γ_2^o, γ_3^o around $\hat{x}$ and repeating the process. The upper confidence limit can be computed similarly. We illustrate with Example 1.

EXAMPLE 1 (continued). Recalling that

$$f(x,\theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}$$

let us set a confidence interval at θ_1. We have

$$SSE_{full} = 0.03049554 \qquad \text{(from Figure 5a)}$$

By simply reusing the SAS code from Figure 9a and embedding it in a MACRO whose argument γ^o is assigned to the paramter θ_1 we can easily construct the following table from Figure 14a.

γ^o	SSE_{γ^o}	$L(\gamma^o)$
-.052	0.03551086	4.275980
-.051	0.03513419	3.954837
-.050	0.03477221	3.646219
-.001	0.03505883	3.890587
.000	0.03543298	4.209581
.001	0.03582188	4.541151

Then either by hand calculator or by using PROC MATRIX as in Figure 14b one can interpolate from this table to obtain the confidence interval

$$[-0.0518,\ 0.0000320].$$

Figure 14a. Likelihood Ratio Test Confidence Interval Construction Illustrated with Example 1.

SAS Statements:

```
%MACRO SSE(GAMMA);
PROC NLIN DATA=EXAMPLE1 METHOD=GAUSS ITER=50 CONVERGENCE=1.0E-13;
PARMS  T2=1.01567967 T3=-1.11569714 T4=-0.50490286;  T1=&GAMMA;
MODEL Y=T1*X1+T2*X2+T4*EXP(T3*X3);
DER.T2=X2; DER.T3=T4*X3*EXP(T3*X3); DER.T4=EXP(T3*X3);
%MEND SSE;
%SSE(-.052) %SSE(-.051) %SSE(-.050) %SSE(-.001) %SSE(.000) %SSE(.001)
```

Output:

NON-LINEAR LEAST SQUARES ITERATIVE PHASE

DEPENDENT VARIABLE: Y METHOD: GAUSS-NEWTON

ITERATION	T2	T3	T4	RESIDUAL SS
6	1.02862742	-1.08499107	-0.49757910	0.03551086
5	1.02812865	-1.08627326	-0.49786686	0.03513419
5	1.02763014	-1.08754637	-0.49815400	0.03477221
7	1.00345514	-1.14032573	-0.51156098	0.03505883
7	1.00296592	-1.14123442	-0.51182277	0.03543298
7	1.00247682	-1.14213734	-0.51208415	0.03582188

Fisure 14b. Likelihood Ratio Test Confidence Interval Construction Illustrated with Example 1.

SAS Statements:

```
PROC MATRIX;
A= 1 -.052 .002704 /
   1 -.051 .002601 /
   1 -.050 .002500 ;
TEST= 4.275980 / 3.954837 / 3.646219 ;  B=INV(A)*TEST;
ROOT=(-B(2,1)+SQRT(B(2,1)#B(2,1)-4#B(3,1)#(B(1,1)-4.22)))#/(2#B(3,1));
PRINT ROOT;
ROOT=(-B(2,1)-SQRT(B(2,1)#B(2,1)-4#B(3,1)#(B(1,1)-4.22)))#/(2#B(3,1));
PRINT ROOT;
A= 1 -.001 .000001 /
   1  .000 .000000 /
   1  .001 .000001 ;
TEST= 3.890587 / 4.209581 / 4.541151 ; B =INV(A)*TEST;
ROOT=(-B(2,1)+SQRT(B(2,1)#B(2,1)-4#B(3,1)#(B(1,1)-4.22)))#/(2#B(3,1));
PRINT ROOT;
ROOT=(-B(2,1)-SQRT(B(2,1)#B(2,1)-4#B(3,1)#(B(1,1)-4.22)))#/(2#B(3,1));
PRINT ROOT;
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

ROOT	COL1
ROW1	0.000108776

ROOT	COL1
ROW1	-0.0518285

ROOT	COL1
ROW1	.0000320109

ROOT	COL1
ROW1	-0.0517626

Next let us set a confidence interval on the parametric function

$$\gamma(\theta) = (\partial/\partial x_3) f(x,\theta)\Big|_{x_3=1} = \theta_3\theta_4 e^{\theta_3}.$$

As we have seen previously, the hypothesis

$$H: \theta_3\theta_4 e^{\theta_3} = \gamma^o$$

can be rewritten as

$$H: \theta_4 = (\theta_3 e^{\theta_3}/\gamma^o)^{-1}.$$

Again, as we have seen previously, to compute SSE_{γ^o} let

$$g_{\gamma^o}(\rho) = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ (\rho_3 e^{\rho_3}/\gamma^o) \end{pmatrix}$$

and SSE_{γ^o} can be computed as the unconstrained minimum of $SSE[q_{\gamma^o}(\rho)]$. Using the SAS code from Figure 9b and embedding it in a MACRO whose argument γ^o replaces the value 1/5 in the previous code the following table can be constructed from Figure 14c.

Figure 14c. Likelihood Ratio Test Confidence Interval Construction Illustrated with Example 1.

SAS Statements:

```
%MACRO SSE(GAMMA);
PROC NLIN DATA=EXAMPLE1 METHOD=GAUSS ITER=60 CONVERGENCE=1.0E-8;
PARMS R1=-0.02588970 R2=1.01567967 R3=-1.11569714;  RG=1/&GAMMA;
T1=R1; T2=R2; T3=R3; T4=1/(RG*R3*EXP(R3));
MODEL Y=T1*X1+T2*X2+T4*EXP(T3*X3);
DER_T1=X1; DER_T2=X2; DER_T3=T4*X3*EXP(T3*X3); DER_T4=EXP(T3*X3);
DER.R1=DER_T1; DER.R2=DER_T2;
DER.R3=DER_T3+DER_T4*(-T4**2)*(RG*EXP(R3)+RG*R3*EXP(R3));
%MEND SSE;
%SSE(.166) %SSE(.167) %SSE(.168) %SSE(.200) %SSE(.201) %SSE(.202)
```

Output:

NON-LINEAR LEAST SQUARES ITERATIVE PHASE

DEPENDENT VARIABLE: Y METHOD: GAUSS-NEWTON

ITERATION	R1	R2	R3	RESIDUAL SS
6	-0.03002338	1.01672014	-0.91765508	0.03591352
8	-0.02978174	1.01642383	-0.93080113	0.03540285
8	-0.02954071	1.01614385	-0.94412575	0.03491101
31	-0.02301828	1.01965423	-1.16048699	0.03493222
43	-0.02283734	1.01994671	-1.16201915	0.03553200
13	-0.02265799	1.02024775	-1.16319256	0.03617013

Figure 14d. Likelihood Ratio Test Confidence Interval Construction Illustrated with Example 1.

SAS Statements:

```
PROC MATRIX;
A= 1 .166 .027556 /
   1 .167 .027889 /
   1 .168 .028224 ;
TEST= 4.619281 / 4.183892 / 3.764558 ;  B=INV(A)*TEST;
ROOT=(-B(2,1)+SQRT(B(2,1)#B(2,1)-4#B(3,1)#(B(1,1)-4.22)))#/(2#B(3,1));
PRINT ROOT;
ROOT=(-B(2,1)-SQRT(B(2,1)#B(2,1)-4#B(3,1)#(B(1,1)-4.22)))#/(2#B(3,1));
PRINT ROOT;
A= 1 .200 .040000 /
   1 .201 .040401 /
   1 .202 .040804 ;
TEST= 3.782641 / 4.294004 / 4.838063 ; B =INV(A)*TEST;
ROOT=(-B(2,1)+SQRT(B(2,1)#B(2,1)-4#B(3,1)#(B(1,1)-4.22)))#/(2#B(3,1));
PRINT ROOT;
ROOT=(-B(2,1)-SQRT(B(2,1)#B(2,1)-4#B(3,1)#(B(1,1)-4.22)))#/(2#B(3,1));
PRINT ROOT;
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

ROOT	COL1
ROW1	0.220322

ROOT	COL1
ROW1	0.166916

ROOT	COL1
ROW1	0.200859

ROOT	COL1
ROW1	0.168861

γ^o	SSE_{γ^o}	$L(\gamma^o)$
.166	0.03591352	4.619281
.167	0.03540285	4.183892
.168	0.03491101	3.764558
.200	0.03493222	3.782641
.201	0.03553200	4.294004
.202	0.03617013	4.838063

Quadratic interpolation from this table as shown in Figure 14d yields

$$[0.1669,\ 0.2009].$$ I

To construct a confidence interval for $\gamma(\theta)$ by inverting the Lagrange multiplier tests, let

$$h(\theta) = \gamma(\theta) - \gamma^o$$

$\tilde{\theta}$ minimize $SSE(\theta)$ subject to $h(\theta) = 0$

$$\tilde{F} = F(\tilde{\theta}) = (\partial/\partial\theta')f(\tilde{\theta})$$

$$\tilde{D} = (\tilde{F}'\tilde{F})^{-1}\tilde{F}'[y - f(\tilde{\theta})]$$

$$R_1(\gamma^o) = [\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D}/q]/[SSE(\hat{\theta})/(n-p)]$$

$$R_2(\gamma^o) = n\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D}/SSE(\tilde{\theta}).$$

The first version of the Lagrange multiplier test accepts when

$$R_1(\gamma^o) < F_\alpha$$

and the second when

$$R_2(\gamma^o) < d_\alpha$$

where $F_\alpha = F^{-1}(1-\alpha;\ q,\ n-p)$, $d_\alpha = nF_\alpha/[(n-p)/q + F_\alpha]$, and q is the dimension of γ^o. Confidence regions consist of those points γ^o for which the tests accept. These confidence regions have the same structural characteristics as likelihood ratio confidence regions except that disjoint islands are much more likely with Lagrange multiplier regions (Problem 2).

In the univariate case, Lagrange multiplier tests are inverted the same as the likelihood ratio test. One constructs a table with $R_1(\gamma^o)$ and $R_2(\gamma^o)$ evaluated at three points around each of the Wald test confidence limits and then uses quadratic interpolation to find the limits. We illustrate with Example 1.

EXAMPLE 1 (continued). Recalling that

$$f(x,\theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}$$

let us set Lagrange multiplier confidence intervals on θ_1. We have

$$SSE(\hat{\theta}) = 0.03049554 \qquad \text{(from Figure 5a)}$$

Taking $\tilde{\theta}$ and $SSE(\tilde{\theta})$ from Figure 14a and embedding the SAS code from Figure 12a

Figure 15a. Lagrange Multiplier Test Confidence Interval Construction Illustrated with Example 1.

SAS Statements:

```
%MACRO DFFD(THETA1,THETA2,THETA3,THETA4,SSER);
DATA WORK01;  SET EXAMPLE1;
T1=&THETA1;  T2=&THETA2;  T3=&THETA3;  T4=&THETA4;
E=Y-(T1*X1+T2*X2+T4*EXP(T3*X3));
F1=X1;  F2=X2;  F3=T4*X3*EXP(T3*X3);  F4=EXP(T3*X3);
DROP T1 T2 T3 T4;
PROC REG DATA=WORK01;  MODEL E=F1 F2 F3 F4 / NOINT;
%MEND DFFD;
%DFFD(-.052, 1.02862742, -1.08499107, -0.49757910, 0.03551086)
%DFFD(-.051, 1.02812865, -1.08627326, -0.49786686, 0.03513419)
%DFFD(-.050, 1.02763014, -1.08754637, -0.49815400, 0.03477221)
%DFFD(-.001, 1.00345514, -1.14032573, -0.51156098, 0.03505883)
%DFFD( .000, 1.00296592, -1.14123442, -0.51182277, 0.03543298)
%DFFD( .001, 1.00247682, -1.14213734, -0.51208415, 0.03582188)
```

Output:

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	4	0.005017024	0.001254256	1.069	0.3916
ERROR	26	0.030494	0.00117284		
U TOTAL	30	0.035511			
MODEL	4	0.004640212	0.001160053	0.989	0.4309
ERROR	26	0.030494	0.001172845		
U TOTAL	30	0.035134			
MODEL	4	0.004278098	0.001069524	0.912	0.4717
ERROR	26	0.030494	0.001172851		
U TOTAL	30	0.034772			
MODEL	4	0.004564169	0.001141042	0.973	0.4392
ERROR	26	0.030495	0.001172871		
U TOTAL	30	0.035059			
MODEL	4	0.004938382	0.001234596	1.053	0.3996
ERROR	26	0.030495	0.001172869		
U TOTAL	30	0.035433			
MODEL	4	0.005327344	0.001331836	1.136	0.3617
ERROR	26	0.030495	0.001172867		
U TOTAL	30	0.035822			

in a MACRO as shown in Figure 15a we obtain the following table from the entries in Figure 15a:

γ^o	$\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D}$	$R_1(\gamma^o)$	$R_2(\gamma^o)$
-.052	0.005017024	4.277433	4.238442
-.051	0.004640212	3.956169	3.962134
-.050	0.004278098	3.647437	3.690963
-.001	0.004564169	3.891336	3.905580
.000	0.004938382	4.210384	4.181174
.001	0.005327344	4.542001	4.461528

Interpolating as shown in Figure 15b we obtain

$$R_1: [-0.0518, 0.0000345]$$

$$R_2: [-0.0518, 0.0000317]$$

In exactly the same way we construct the following table for

$$\gamma(\theta) = \theta_3\theta_4 e^{\theta_3}$$

from the entries of Figures 14c and 15c.

Figure 15b. Lagrange Multiplier Test Confidence Interval Construction Illustrated with Example 1.

SAS Statements:

```
PROC MATRIX;
A= 1 -.052 .002704 /
   1 -.051 .002601 /
   1 -.050 .002500 ;
TEST= 4.277433 4.238442 /
      3.956169 3.962134 /
      3.647437 3.690963 ;  B=INV(A)*TEST;
ROOT1=(-B(2,1)+SQRT(B(2,1)#B(2,1)-4#B(3,1)#(B(1,1)-4.22)))#/(2#B(3,1));
ROOT2=(-B(2,1)-SQRT(B(2,1)#B(2,1)-4#B(3,1)#(B(1,1)-4.22)))#/(2#B(3,1));
ROOT3=(-B(2,2)+SQRT(B(2,2)#B(2,2)-4#B(3,2)#(B(1,2)-4.19)))#/(2#B(3,2));
ROOT4=(-B(2,2)-SQRT(B(2,2)#B(2,2)-4#B(3,2)#(B(1,2)-4.19)))#/(2#B(3,2));
PRINT ROOT1 ROOT2 ROOT3 ROOT4;
A= 1 -.001 .000001 /
   1  .000 .000000 /
   1  .001 .000001 ;
TEST= 3.891336 3.905580 /
      4.210384 4.181174 /
      4.452001 4.461528 ;  B=INV(A)*TEST;
ROOT1=(-B(2,1)+SQRT(B(2,1)#B(2,1)-4#B(3,1)#(B(1,1)-4.22)))#/(2#B(3,1));
ROOT2=(-B(2,1)-SQRT(B(2,1)#B(2,1)-4#B(3,1)#(B(1,1)-4.22)))#/(2#B(3,1));
ROOT3=(-B(2,2)+SQRT(B(2,2)#B(2,2)-4#B(3,2)#(B(1,2)-4.19)))#/(2#B(3,2));
ROOT4=(-B(2,2)-SQRT(B(2,2)#B(2,2)-4#B(3,2)#(B(1,2)-4.19)))#/(2#B(3,2));
PRINT ROOT1 ROOT2 ROOT3 ROOT4;
```

Output:

```
            S T A T I S T I C A L   A N A L Y S I S   S Y S T E M          1

                    ROOT1          COL1
                    ROW1      .0000950422

                    ROOT2          COL1
                    ROW1       -0.0518241

                    ROOT3          COL1
                    ROW1        0.0564016

                    ROOT4          COL1
                    ROW1       -0.051826

                    ROOT1          COL1
                    ROW1      .0000344662

                    ROOT2          COL1
                    ROW1       0.00720637

                    ROOT3          COL1
                    ROW1      .0000317425

                    ROOT4          COL1
                    ROW1       -0.116828
```

Figure 15c. Lagrange Multiplier Test Confidence Interval Construction Illustrated with Example 1.

SAS Statements:

```
%MACRO DFFD(GAMMA,RHO1,RHO2,RHO3,SSER);
DATA WORK01;  SET EXAMPLE1;
T1=&RHO1;  T2=&RHO2;  T3=&RHO3;  T4=1/(&RHO3*EXP(&RHO3)/&GAMMA);
E=Y-(T1*X1+T2*X2+T4*EXP(T3*X3));
F1=X1;  F2=X2;  F3=T4*X3*EXP(T3*X3);  F4=EXP(T3*X3);
DROP T1 T2 T3 T4;
PROC REG DATA=WORK01;  MODEL E=F1 F2 F3 F4 / NOINT;
%MEND DFFD;
%DFFD( .166, -0.03002338, 1.01672014, -0.91765508, 0.03591352)
%DFFD( .167, -0.02978174, 1.01642383, -0.93080113, 0.03540285)
%DFFD( .168, -0.02954071, 1.01614385, -0.94412575, 0.03491101)
%DFFD( .200, -0.02301828, 1.01965423, -1.16048699, 0.03493222)
%DFFD( .201, -0.02283734, 1.01994671, -1.16201915, 0.03553200)
%DFFD( .202, -0.02265799, 1.02024775, -1.16319256, 0.03617013)
```

Output:

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	4	0.005507692	0.001376923	1.177	0.3438
ERROR	26	0.030406	0.001169455		
U TOTAL	30	0.035914			
MODEL	4	0.004986108	0.001246527	1.066	0.3935
ERROR	26	0.030417	0.001169875		
U TOTAL	30	0.035403			
MODEL	4	0.004483469	0.001120867	0.958	0.4471
ERROR	26	0.030428	0.00117029		
U TOTAL	30	0.034911			
MODEL	4	0.004439308	0.001109827	0.946	0.4531
ERROR	26	0.030493	0.001172804		
U TOTAL	30	0.034932			
MODEL	4	0.005039249	0.001259812	1.074	0.3894
ERROR	26	0.030493	0.001172798		
U TOTAL	30	0.035532			
MODEL	4	0.005677511	0.001419378	1.210	0.3303
ERROR	26	0.030493	0.001172793		
U TOTAL	30	0.036170			

γ^o	$D'(F'F)D$	$R_1(\gamma^o)$	$R_2(\gamma^o)$
.166	0.005507692	4.695768	4.600795
.167	0.004986108	4.251074	4.225175
.168	0.004483469	3.822533	3.852770
.200	0.004439308	3.784882	3.812504
.201	0.005039249	4.296382	4.254685
.202	0.005677511	4.840553	4.709005

Quadratic interpolation from this table as shown in Figure 15d yields

R_1: [0.1671, 0.2009]

R_2: [0.1671, 0.2009] ∎

There is some risk in using quadratic interpolation around Wald test confidence limits to find likelihood ratio or Lagrange multiplier confidence intervals. If the confidence region is a union of disjoint intervals then the method will compute the wrong answer. To be completely safe one would have to plot $L(\gamma^o)$, $R_1(\gamma^o)$, or $R_2(\gamma^o)$ and inspect for local minima.

The usual criterion for judging the quality of a confidence procedure is expected length, area, or volume depending on the dimension q of $\gamma(\theta)$. Let us use volume as the generic term. If two confidence procedures have the same probability of covering $\gamma(\theta^o)$ then the one with the smallest expected volume is preferred. But expected volume is really just an attribute of the power curve of the test to which the confidence procedure corresponds. To see this, let a test be described by its critical function

Figure 15d. Lagrange Multiplier Test Confidence Interval Construction Illustrated with Example 1.

SAS Statements:

```
PROC MATRIX;
A= 1 .166 .027556 /
   1 .167 .027889 /
   1 .168 .028224 ;
TEST= 4.695768 4.600795 /
      4.251074 4.225175 /
      3.822533 3.852770 ;  B=INV(A)*TEST;
ROOT1=(-B(2,1)+SQRT(B(2,1)#B(2,1)-4#B(3,1)#(B(1,1)-4.22)))#/(2#B(3,1));
ROOT2=(-B(2,1)-SQRT(B(2,1)#B(2,1)-4#B(3,1)#(B(1,1)-4.22)))#/(2#B(3,1));
ROOT3=(-B(2,2)+SQRT(B(2,2)#B(2,2)-4#B(3,2)#(B(1,2)-4.19)))#/(2#B(3,2));
ROOT4=(-B(2,2)-SQRT(B(2,2)#B(2,2)-4#B(3,2)#(B(1,2)-4.19)))#/(2#B(3,2));
PRINT ROOT1 ROOT2 ROOT3 ROOT4;
A= 1 .200 .040000 /
   1 .201 .040401 /
   1 .202 .040804 ;
TEST= 3.784882 3.812504 /
      4.296382 4.254685 /
      4.840553 4.709005 ;  B=INV(A)*TEST;
ROOT1=(-B(2,1)+SQRT(B(2,1)#B(2,1)-4#B(3,1)#(B(1,1)-4.22)))#/(2#B(3,1));
ROOT2=(-B(2,1)-SQRT(B(2,1)#B(2,1)-4#B(3,1)#(B(1,1)-4.22)))#/(2#B(3,1));
ROOT3=(-B(2,2)+SQRT(B(2,2)#B(2,2)-4#B(3,2)#(B(1,2)-4.19)))#/(2#B(3,2));
ROOT4=(-B(2,2)-SQRT(B(2,2)#B(2,2)-4#B(3,2)#(B(1,2)-4.19)))#/(2#B(3,2));
PRINT ROOT1 ROOT2 ROOT3 ROOT4;
```

Output:

S T A T I S T I C A L A N A L Y S I S S Y S T E M 1

ROOT1	COL1
ROW1	0.220989

ROOT2	COL1
ROW1	0.167071

ROOT3	COL1
ROW1	0.399573

ROOT4	COL1
ROW1	0.167094

ROOT1	COL1
ROW1	0.200855

ROOT2	COL1
ROW1	0.168833

ROOT3	COL1
ROW1	0.200855

ROOT4	COL1
ROW1	0.127292

$$\phi(y,\gamma^o) = \begin{cases} 1 & \text{reject H: } \gamma(\theta) = \gamma^o \\ 0 & \text{accept H: } \gamma(\theta) = \gamma^o . \end{cases}$$

The corresponding confidence procedure is

$$R_y = \{\gamma_o : \phi(y,\gamma_o) = 0\}.$$

Expected volume is computed as

$$\text{Expected volume } (\phi) = \int_{R^n} \int_{R_y} d\gamma dN[y;\ f(\theta^o),\ \sigma^2 I]$$

$$= \int_{R^n} \int_{R^q} [1 - \phi(y,\gamma)]d\gamma dN[y;\ f(\theta^o),\ \sigma^2 I]$$

As Pratt (1961) shows by interchanging the order of integration

$$\text{Expected volume } (\phi) = \int_{R^q} \int_{R^n} [1 - \phi(y,\gamma)]dN(y;\ f(\theta^o)\ ,\ \sigma^2 I]d\gamma$$

$$= \int_{R^q} P[\phi(y,\gamma) = 0 \mid \theta^o,\ \sigma^2]d\gamma$$

The integrand is the probability of covering γ,

$$c_\phi(\gamma) = P[\phi(y,\gamma) = 0 \mid \theta^o,\ \sigma^2]$$

and is analogous to the operating characteristic curve of a test. The essential difference between the coverage function $c_\phi(\gamma)$ and the operating characteristic function lies in the treatment of the hypothesized value γ and the true value of the parameter θ^o. For the coverage function, θ^o is held

fixed and γ varies; the converse is true for the operating characteristic function. If a test $\phi(y,\gamma)$ has better power against H: $\gamma(\theta) = \gamma^o$ than the test $\psi(y,\gamma^o)$ for all γ^o then we have that

$$c_\phi(\gamma^o) = P[\phi(y,\gamma^o) = 0 \mid \theta^o, \sigma^2]$$

$$\leq P[\psi(y,\gamma^o) = 0 \mid \theta^o, \sigma^2]$$

$$= c_\psi(\gamma^o)$$

which implies

$$\text{Expected volume } (\phi) \leq \text{Expected volume } (\psi).$$

In this case a confidence procedure based on ϕ is to be preferred to a confidence interval based on ψ.

If one accepts the approximations of the previous section as giving useful guidance in applications then the confidence procedure obtained by inverting the likelihood ratio test is to be preferred to either of the Lagrange multiplier procedures. However, both the likelihood ratio and Lagrange procedures can have infinite expected volume; Example 2 is an instance (Problem 3). But for $\gamma \neq \gamma(\theta^o)$ the coverage function gives the probability that the confidence procedure covers false values of γ. Thus, even in the case of infinite expected volume, the inequality $c_\phi(\gamma) \leq c_\psi(\gamma)$ implies that the procedure obtained by inverting ϕ is preferred to that obtained by inverting ψ. Thus the likelihood ratio procedure remains preferable to the Lagrange multiplier procedures even in the case of infinite expected volume.

Again, if one accepts the approximations of the previous section, the confidence procedure obtained by inverting the Wald test has better structural characteristics than either the likelihood ratio procedure or the Lagrange multiplier procedures. Wald test confidence regions are always intervals, ellipses, or ellipsoids according to the dimension of $\gamma(\theta)$ and they are much easier to compute than likelihood ratio or Lagrange multiplier regions. Expected volume is always finite (Problem 4). It is a pity that the accuracy of the approximation to the probability $P(W > F_\alpha)$ by $P(Y > F_\alpha)$ of the previous section is often inaccurate. This makes use of Wald confidence regions risky as one cannot be sure that the actual coverage probability is accurately approximated by the nominal probability of $1-\alpha$ short of Monte Carlo simulation at each instance. In the next chapter we shall consider methods that are intended to remedy this defect.

PROBLEMS

1. In the notation of the last few paragraphs of this section show that

$$P\{\phi[y, \gamma(\theta^o)] = 0 \mid \theta^o, \sigma\} = \int_{R_y} dN[y; f(\theta^o), \sigma^2 I].$$

2. (Disconnected confidence regions.) Fill in the missing details in the following argument. Consider setting a confidence region on the entire parameter vector θ. Islands in likelihood ratio confidence regions may occur because $SSE(\theta)$ has a local minimum at θ^* causing $L(\theta^*)$ to fall below F_α. But if θ^* is a local minimum then $R_1(\theta^*) = R_2(\theta^*) = 0$ and a neighborhood of θ^* must be included in a Lagrange multiplier confidence region.

3. Referring to Model B of Example 2 and the hypothesis H: $\theta^o = \gamma^o$ show that the fact that $0 < f(x,\gamma) < 1$ implies that $P(X > c_\alpha) < 1$ for all γ in $A = \{\gamma: \ 0 \leq \gamma_2 \leq \gamma_1\}$ where X and c_α are as defined in the previous section. Show also that there is an open set E such that for all e in E we have

$$\sup_{\gamma \in A} \|e + \delta(\gamma)\| \leq c_\alpha \inf_{\gamma \in A} \|e + \delta(\gamma)\|^2$$

where $\delta(\gamma) = f(\theta^o) - f(\gamma)$. Show that this implies that $P(L > F_\alpha) < 1$ for all γ in A. Show that these facts imply that the expected volume of the likelihood ratio confidence region is infinite both when the approximating random variable X is used in the computation and when L itself is used.

4. Show that if $Y \sim F'[q, n-p, \lambda(\gamma^o)]$ where

$$\lambda(\gamma^o) = [\gamma(\theta^o) - \gamma^o]'\{H(\theta^o)[F'(\theta^o)F(\theta^o)]^{-1}H'(\theta^o)\}^{-1}[\gamma(\theta^o) - \gamma^o]/(2\alpha)^2$$

and $c_Y(\gamma^o) = P(Y \leq F_\alpha)$ then $\int_{R^q} c_Y(\gamma)d\gamma < \infty$.

7. REFERENCES

Bartle, Robert G. (1964), The Elements of Real Analysis. New York: John Wiley and Sons.

Beale, E. M. L. (1960), "Confidence Regions in Non-Linear Estimation," Journal of the Royal Statistical Society, Series B, 22, 41-76.

Blackwell, D. and M. A. Girshick (1954), Theory of Games and Statistical Decisions. New York: John Wiley and Sons.

Box, G. E. P. and H. L. Lucus (1959), "The Design of Experiments in Non-Linear Situations," Biometrika 46, 77-90.

Dennis, J. E., D. M. Gay and Roy E. Welch (1977), "An Adaptive Nonlinear Least-Squares Algorithm," Department of Computer Sciences Report No. TR 77-321, Cornell University, Ithaca, New York.

Fox, M. (1956), "Charts on the Power of the T-Test," The Annals of Mathematical Statistics 27, 484-497.

Gallant, A. Ronald (1973), "Inference for Nonlinear Models," Institute of Statistics Mimeograph Series No. 875, North Carolina State University, Raleigh, North Carolina.

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 (1976), "Confidence Regions for the Parameters of a Nonlinear Regression Model," Institute of Statistics Mimeograph Series No. 875, North Carolina State University, Raleigh, North Carolina.

Gallant, A. Ronald (1980), "Explicit Estimators of Parametric Functions in Nonlinear Regression," Journal of the American Statistical Association 75, 182-193.

Gill, Philip E., Walter Murray and Margaret H. Wright (1981), Practical Optimization. New York: Academic Press.

Golub, Gene H. and Victor Pereyra (1973), "The Differentiation of Psuedo-Inverses and Nonlinear Least-Squares Problems whose Variable Separate," SIAM Journal of Numerical Analysis 10, 413-432.

Guttman, Irwin and Duane A. Meeter (1964), "On Beale's Measures of Non-Linearity," Technometrics 7, 623-637.

Hammersley, J. M. and D. C. Handscomb (1964), Monte Carlo Methods. New York: John Wiley and Sons.

Hartley, H. O. (1961), "The Modified Gauss-Newton Method for the Fitting of Nonlinear Regression Functions by Least Squares," Technometrics 3, 269-280.

Hartley, H. O. and A. Booker (1965), "Nonlinear Least Squares Estimation," Annals of Mathematical Statistics 36, 638-650.

Huber, Peter (1982), "Comment on the Unification of the Asymptotic Theory of Nonlinear Econometric Models," Econometric Reviews 1, 191-192.

Jensen, D. R. (1981), "Power of Invariant Tests for Linear Hypotheses under Spherical Symmetry," Scandanavian Journal of Statistics 8, 169-174.

Levenberg, K. (1944), "A Method for the Solution of Certain Problems in Least Squares," Quarterly Journal of Applied Mathematics 2, 164-168.

Malinvaud, E. (1970), Statistical Methods of Econometrics (Chapter 9). Amsterdam: North-Holland.

Marquardt, Donald W. (1963), "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," Journal of the Society for Industrial and Applied Mathematics 11, 431-441.

Osborne, M. R. (1972), "Some Aspects of Non-Linear Least Squares Calculations," in Lootsma, F. A. (ed.), Numerical Methods for Non-Linear Optimization. New York: Academic Press.

Pearson, E. S and H. O. Hartley (1951), "Charts of the Power Function of the Analysis of Variance Tests, Derived from the Non-Central F-Distribution," Biometrika 38, 112-130.

Pratt, John W. (1961), "Length of Confidence Intervals," Journal of the American Statistical Association 56, 549-567.

Royden, H. L. (1963), Real Analysis. New York: MacMillan Company.

Scheffé, Henry (1959), The Analysis of Variance. New York: John Wiley and Sons.

Searle, S. R. (1971), Linear Models. New York: John Wiley and Sons.

Tucker, Howard G. (1967), A Graduate Course in Probability. New York: Academic Press.

Zellner, Arnold (1976), "Bayesian and Non-Bayesian Analysis of the Regression Model with Multivariate Student-t Error Terms," Journal of the American Statistical Association 71, 400-405.

8. INDEX TO CHAPTER 1.

Chain rule, 1-2-3, 1-2-11
Compartment analysis, 1-1-7
Composite function rule, 1-2-3, 1-2-11
Confidence regions
    correspondence between expected length, area, or volume and power of a test, 1-6-21
    Lagrange multiplier, 1-6-14
    likelihood ratio, 1-6-6
    structural characteristics of, 1-6-6, 1-6-14, 1-6-22, 1-6-24
    Wald, 1-6-1
Coverage function, 1-6-21
Critical function, 1-6-21
Differentiation
    chain rule, 1-2-3, 1-2-11
    composite function rule, 1-2-3, 1-2-11
    gradient, 1-2-1
    Jacobian, 1-2-2
    hessian, 1-2-1
    matix derivative, 1-2-1
    vector derivative, 1-2-1
Disconnected confidence regions, 1-6-24
Efficient score test
    (see Lagrange multiplier test)
Figure 1, 1-4-2
Figure 2, 1-4-3
Figure 3, 1-4-9
Figure 4, 1-4-12
Figure 5a, 1-4-14
Figure 5b, 1-4-16
Figure 6, 1-4-22
Figure 7, 1-5-8
Figure 8, 1-5-10
Figure 9a, 1-5-20
Figure 9b, 1-5-24
Figure 9c, 1-5-26
Figure 10a, 1-5-29
Figure 10b, 1-5-30
Figure 11a, 1-5-43
Figure 11b, 1-5-45
Figure 11c, 1-5-46
Figure 12a, 1-5-63
Figure 12b, 1-5-66
Figure 12c, 1-5-70
Figure 13, 1-6-4
Figure 14a, 1-6-8
Figure 14b, 1-6-9
Figure 14c, 1-6-11
Figure 14d, 1-6-12
Figure 15a, 1-6-15
Figure 15b, 1-6-17
Figure 15c, 1-6-18
Figure 15d, 1-6-20

Functional dependency, 1-5-16
Gauss-Newton method
    algorithm, 1-4-4
    algorithm failure, 1-4-21
    convergence proof, 1-4-27
    informal discussion, 1-4-1
    starting values, 1-4-6
    step length determination, 1-4-5
    stopping rules, 1-4-5
Gradient, 1-2-1
Grid search, 1-4-17
Jacobian, 1-2-2
Hartley's method
    (see Gauss-Newton method)
Hessian, 1-2-1
Identification Condition, 1-3-7
Lagrange multiplier test
    asymptotic distribution, 1-5-57
    computation, 1-5-62
    corresponding confidence region, 1-6-14
    defined, 1-5-61
    informal discussion, 1-5-55, 1-5-81
    Monti Carlo simulations, 1-5-77
    power computations, 1-5-72
Large residual problem, 1-4-21
Least squares estimator
    characterized as a linear function of the errors, 1-3-1
    computation
        (see Gauss-Newton, Levenberg-Marquardt, and Newton methods)
    defined, 1-2-10
    distribution of, 1-3-2, 1-3-3
    first order conditions, 1-2-2
    informal discussion of regularity conditions, 1-3-5
Least squares scale estimator
    characterized as a quadratic function of the errors, 1-3-2
    computation
        (see Gauss-Newton, Levenberg-Marquardt, and Newton methods)
    defined, 1-2-10
    distribution of, 1-3-2, 1-3-3
Likelihood ratio test
    asymptotic distribution, 1-5-35
    computation, 1-5-16
    corresponding confidence region, 1-6-6
    defined, 1-5-15
    informal discussion, 1-5-13
    Monti Carlo simulations, 1-5-49, 1-5-51, 1-5-54
    power computations, 1-5-32
Linear regression model
    (see univariate nonlinear regression model)
Marquardt's method
    (see Levenberg-Marquardt method)
Matrix derivatives, 1-2-1

Modified Gauss-Newton method
(see Gauss-Newton method)
Nonlinear regression model
(see univariate nonlinear regression model)
Parametric restriction, 1-5-16
Rank Condition, 1-3-7
Rao's efficient score test
(see Lagrange multiplier test)
Table 1, 1-1-5
Table 2, 1-1-9
Table 3, 1-3-13
Table 4, 1-4-24
Table 5, 1-5-12
Table 6, 1-5-36
Table 7, 1-5-48
Table 8, 1-5-50
Table 9, 1-5-52
Table 10a, 1-5-76
Table 10b, 1-5-78
Taylor's theorem, 1-2-8
Univariate linear regression model
defined, 1-1-2
Univariate nonlinear regression model
defined, 1-1-1, 1-1-3
vector representation, 1-2-4
Vector derivatives, 1-2-1
Wald test
asymptotic distribution, 1-5-7
corresponding confidence region, 1-6-1
defined, 1-5-3
informal discussion, 1-5-1
Monti Carlo simulations, 1-3-14, 1-5-13, 1-5-54
power computations, 1-5-9