Nonlinear Statistical Models

A. RONALD GALLANT

Professor of Statistics and Economics
North Carolina State University
Raleigh, North Carolina

JOHN WILEY & SONS

New York • Chichester • Brisbane • Toronto • Singapore

This Page Intentionally Left Blank

Nonlinear Statistical Models

Nonlinear Statistical Models

A. RONALD GALLANT

Professor of Statistics and Economics
North Carolina State University
Raleigh, North Carolina

JOHN WILEY & SONS

New York • Chichester • Brisbane • Toronto • Singapore

Text Credits

SAS® is the registered trademark of SAS Institute Inc., Cary, North Carolina, USA.

Tables 1, 3, 5, 7, Chapter 1. Reprinted by permission from A. Ronald Gallant, Nonlinear regression, *The American Statistician* **29**, 73–81, © 1975 by the American Statistical Association, Washington, D.C. 20005.

Table 4, Figure 6, Chapter 1. Reprinted by permission from A. Ronald Gallant, Testing a nonlinear regression specification: A nonregular case, *The Journal of The American Statistical Association* **72**, 523–530, © 1977 by the American Statistical Association, Washington, D.C. 20005.

Table 6, 8, Chapter 1. Reprinted by permission from A. Ronald Gallant, The power of the likelihood ratio test of location in a nonlinear regression model, *The Journal of The American Statistical Association* **70**, 199–203, © 1975 by the American Statistical Association, Washington, D.C. 20005.

Table 1, Chapter 2. Courtesy of the authors: A. Ronald Gallant and J. Jeffery Goebel, Nonlinear regression with autocorrelated errors, *The Journal of The American Statistical Association* **71** (1976), 961–967.

Table 2, Figure 4, Chapter 2. Reprinted by permission from A. Ronald Gallant, Testing a nonlinear regression specification: A nonregular case, *The Journal of The American Statistical Association* **72**, 523–530, © 1977 by the American Statistical Association, Washington, D.C. 20005.

Figures 1, 2, Chapter 2. Reprinted by permission from A. Ronald Gallant and J. Jeffery Goebel, Nonlinear regression with autocorrelated errors, *The Journal of The American Statistical Association* **71**, 961–967, © 1976 by the American Statistical Association, Washington, D.C. 20005.

Tables 1*a*, 1*b*, 1*c*, Chapter 5. Courtesy of the authors: A. Ronald Gallant and Roger W. Koenker, Cost and benefits of peak-load pricing of electricity: A continuous-time econometric approach, *Journal of Econometrics* **26** (1984), 83–114.

Table 5, Chapter 5. Reprinted by permission from *Statistical Methods*, Seventh Edition, by George W. Snedecor and William G. Cochran, © 1980 by The Iowa State University Press, Ames, Iowa 50010.

Tables 6, 7, Chapter 5. Reprinted by permission from A. Ronald Gallant, On the bias in flexible functional forms and essentially unbiased form: The fourier flexible form, *The Journal of Econometrics* **15**, 211–245, © 1981 by North Holland Publishing Company, Amsterdam.

Tables 1*a*, 1*b*, Chapter 6. Courtesy of the authors: Lars Peter Hansen and Kenneth J. Singleton, Generalized instrumental variables estimators of nonlinear rational expectations models, *Econometrica* **50** (1982) 1269–1286.

Library of Congress Cataloging in Publication Data:

Gallant, A. Ronald, 1942–
Nonlinear statistical models.

(Wiley series in probability and mathematical
statistics. Applied probability and statistics)
Bibliography: p.
Includes index.
1. Regression analysis. 2. Multivariate analysis.
3. Nonlinear theories. I. Title. II. Series.
QA278.2.G35 1987 519.5'4 86-18955
ISBN 0-471-80260-3

10 9 8 7 6 5 4 3 2

To Marcia, Megan, and Drew

This Page Intentionally Left Blank

Preface

Any type of statistical inquiry in business, government, or academics in which principles from some body of knowledge enter seriously into the analysis is likely to lead to a nonlinear statistical model. For instance, a model obtained as the solution of a differential equation arising in engineering, chemistry, or physics is usually nonlinear. Other examples are economic models of consumer demand or of intertemporal consumption and investment.

Much applied work using linear models represents a distortion of the underlying subject matter. In the past there was little else that one could do, given the restrictions imposed by the cost of computing equipment and the lack of an adequate statistical theory. But the availability of computing resources is no longer a problem, and advances in statistical and probability theory have occurred over the last fifteen years that effectively remove the restriction of inadequate theory.

In this book, I have attempted to bring these advances together in one place, organize them, and relate them to applications, for the use of students as a text and for the use of those engaged in research as a reference. My hopes and goals in writing it will be achieved if it becomes possible for the reader to bring subject matter considerations directly to bear on data without distortion.

The coverage is comprehensive. The three major categories of statistical models relating dependent variables to explanatory variables are covered: univariate regression models, multivariate regression models, and simultaneous equations models. These models can have the classical regression structure where the independent variables are ancillary and the errors independent, or they can be dynamic, with lagged dependent variables permitted as explanatory variables and with serially correlated errors. The coverage is also comprehensive in the sense that the subject is treated at all

levels: methods, theory, and computations. However, only material that I think is of practical value in making a statistical inference using a model that derives from subject matter considerations is included.

The statistical methods are accessible to anyone with a good working knowledge of the theory and methods of linear statistical models as found in a text such as Searle's *Linear Models*. It is important that Chapter 1 be read first. It lays the intuitive foundation. There the subject of univariate nonlinear regression is presented by relying on analogy with the theory and methods of linear models, on examples, and on Monte Carlo simulations. 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. One can get the main ideas across and save the theory for later. Generalized least squares can be applied in nonlinear regression just as in linear regression. Using this as a vehicle, the ideas, intuition, and statistical methods developed in Chapter 1 are extended to other situations, notably multivariate nonlinear regression in Chapter 5 and nonlinear simultaneous equations models in Chapter 6. These chapters include many numerical examples.

Chapter 3 is a unified theory of statistical inference for nonlinear models with regression structure, and Chapter 7 is the same for dynamic models. Some useful specialization of the general theory is possible in the case of the univariate nonlinear regression model, and this is done in Chapter 4. Notation, assumptions, and theorems are isolated and clearly identified in the theoretical chapters so that the results can be reliably applied to new situations without need for a detailed reading of the mathematics. These results should be usable by anyone who is comfortable thinking of a random variable as a function defined on an abstract probability space and understands the notion of almost sure convergence. Aside from that, application of the theory does not rise above an advanced calculus level probability course. There are examples in these chapters to provide templates.

Reading the proofs requires a good understanding of measure theoretic probability theory, as would be imparted by a course out of Tucker's *Graduate Course in Probability*, and a working knowledge of analysis, as in Royden's *Real Analysis*. For the reader's convenience, references are confined to these two books as much as possible, but this material is standard and any similar textbook will serve.

The material in Chapter 7 is at the frontier. This is the first time some of it will appear in print. As with anything new, much improvement is still possible. Regularity conditions are more onerous than need be, and there is a paucity of worked examples to determine which of them most need relaxing. I have included full details in the proofs, and have supplied the

details of proofs that seemed too terse in the original source, in hopes that readers can learn the ideas and methods of proof quickly and will move the field forward.

As to computations, one must either use a programming language, with or without the aid of a scientific subroutine library, or use a statistical package. Hand calculator computations are out of the question. Using a programming language to present the ideas seems ill advised. Discussion bogs down in detail that is just tedious accounting and has nothing to do with the subject proper. For pedagogical purposes, a statistical package is the better choice. Its code should be concise and readable, even to the uninitiated. I chose SAS®, and it seems to have served well. Computational examples consist of figures displaying a few lines of SAS code and the resulting output. For those who would rather use a programming language in applications, the algorithms are in the text, and anyone accustomed to using a programming language should have no trouble implementing them; the examples will be helpful in debugging.

I have debts to acknowledge. The biggest is to my family. Hours—no, years—were spent writing that ought to have been spent with them. I owe a debt to my students Geraldo Souza and Jose Francisco Burguete. The theory for models with regression structure is their dissertation research. The theory for dynamic models was worked out while Halbert White and Jeffrey Wooldridge visited Raleigh in the summer of 1984, and much of it is theirs. I owe a special debt to my secretary, Janice Gaddy. She typed the manuscript cheerfully, promptly, and accurately. More importantly, she held every annoyance at bay.

Support while writing this book was provided by National Science Foundation Grants SES 82-07362 and SES 85-07829, North Carolina Agricultural Experiment Station Projects NC03641, NC03879, and NC05593, and the PAMS Foundation. SAS Institute Inc. let me use its computing equipment and a prerelease version of PROC SYSNLIN for the computations in Chapter 6 and has, over the years, provided generous support to the Triangle Econometrics Workshop. Many ideas in this book have come from that workshop.

A. RONALD GALLANT

December, 1986
Raleigh, North Carolina

This Page Intentionally Left Blank

Contents

Nonlinear Statistical Models

This Page Intentionally Left Blank

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 because 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.

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

$$y_t = f(x_t, \theta^0) + e_t \qquad t = 1, 2, \ldots, n$$

where $f(x, \theta)$ is the known response function, θ^0 is a p-dimensional vector of unknown parameters, and the e_t represent unobservable observational or experimental errors. We write θ^0 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 Chapter 3 for additional details on this point, and Section 8 of Chapter 3 for a device that allows one to consider the random regressor setup 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.

$$\varphi_0(y_t) = \sum_{i=1}^{p} \varphi_i(x_t)\theta_i + e_t$$

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

$$\varphi_0(y_t) = \sum_{i=1}^{p} \varphi_i(x_t)\theta_i + e_t.$$

A univariate nonlinear regression model is of the form

$$\varphi_0(y_t) = f(x_t, \theta) + e_t$$

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 noncentral chi-square distribution. This is somewhat 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

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

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	5.99
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	0.97656	0	1	2.65
29	0.50811	1	1	0.08
30	0.91840	0	1	6.11

Source: Gallant (1975d).

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

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

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

$$f(x_1, \theta^0) = \theta_1^0 + \theta_2^0 + \theta_4^0 e^{6.28\theta_3^0}.$$

Similarly, the second observation

$$x_2 = (0, 1, 9.86)'$$

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

$$f(x_2, \theta^0) = \theta_2^0 + \theta_4^0 e^{9.86\theta_3^0}$$

and so on. The parameter θ_1^0 is 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, 1965; Gallant, 1980) that it might be called the standard pedagogical example.

Linear System

$$\frac{d}{dx}A(x) = -\theta_1 A(x)$$

$$\frac{d}{dx}B(x) = \theta_1 A(x) - \theta_2 B(x)$$

$$\frac{d}{dx}C(x) = \theta_2 B(x)$$

Boundary Conditions

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

Parameter Space

$$\theta_1 \geq \theta_2 \geq 0$$

Solution, $\theta_1 > \theta_2$

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

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

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

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 (1965), we shall use the solutions for $B(x)$ and $C(x)$ to construct two nonlinear models (see Table 2) 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 (1965).

Model B

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

$$\theta^0 = (1.4, .4)'$$

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

$$n = 12$$

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

Model C

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

$$\theta^0 = (1.4, .4)'$$

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

$$n = 12$$

$$\sigma^2 = (.025)^2$$ □

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

A word regarding notation. All vectors, such as θ, are column vectors unless the contrary is indicated by θ', which is a row vector. Strict adherence to this convention in notation leads to clutter, such as

$$d = (a', b', c')'.$$

We shall usually let the primes be understood in these cases and write

$$d = (a, b, c)$$

instead. Transposition will be carefully indicated at instances where clarity seems to demand it.

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 $s(\theta)$,

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

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

$$\underset{1}{\frac{\partial}{\partial\theta'}s(\theta)} = \underset{1}{\left(\frac{\partial}{\partial\theta_1}s(\theta), \frac{\partial}{\partial\theta_2}s(\theta), \ldots, \frac{\partial}{\partial\theta_p}s(\theta)\right)}_{p}.$$

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

$s(\theta)$,

$$\frac{\partial^2}{\partial\theta\,\partial\theta'}s(\theta) = {}_p\begin{pmatrix} \frac{\partial^2}{\partial\theta_1^2}s(\theta) & \frac{\partial^2}{\partial\theta_1\,\partial\theta_2}s(\theta) & \cdots & \frac{\partial^2}{\partial\theta_1\,\partial\theta_p}s(\theta) \\ \frac{\partial^2}{\partial\theta_2\,\partial\theta_1}s(\theta) & \frac{\partial^2}{\partial\theta_2^2}s(\theta) & \cdots & \frac{\partial^2}{\partial\theta_2\,\partial\theta_p}s(\theta) \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2}{\partial\theta_p\,\partial\theta_1}s(\theta) & \frac{\partial^2}{\partial\theta_p\,\partial\theta_2}s(\theta) & \cdots & \frac{\partial^2}{\partial^2\theta_p}s(\theta) \end{pmatrix}_p .$$

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

$$f(\theta) = {}_n\begin{pmatrix} f_1(\theta) \\ f_2(\theta) \\ \vdots \\ f_n(\theta) \end{pmatrix}_1$$

is the n by p matrix

$$\frac{\partial}{\partial\theta'}f(\theta) = {}_n\begin{pmatrix} \frac{\partial}{\partial\theta_1}f_1(\theta) & \frac{\partial}{\partial\theta_2}f_1(\theta) & \cdots & \frac{\partial}{\partial\theta_p}f_1(\theta) \\ \frac{\partial}{\partial\theta_1}f_2(\theta) & \frac{\partial}{\partial\theta_2}f_2(\theta) & \cdots & \frac{\partial}{\partial\theta_p}f_2(\theta) \\ \vdots & \vdots & & \vdots \\ \frac{\partial}{\partial\theta_1}f_n(\theta) & \frac{\partial}{\partial\theta_2}f_n(\theta) & \cdots & \frac{\partial}{\partial\theta_p}f_n(\theta) \end{pmatrix}_p .$$

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

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

Then

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

In this notation, the following rule governs matrix transposition:

$$\left(\frac{\partial}{\partial\theta'}f(\theta)\right)' = \frac{\partial}{\partial\theta}f'(\theta).$$

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

$$\begin{aligned} \frac{\partial^2}{\partial\theta\,\partial\theta'}s(\theta) &= \frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta'}s(\theta)\right) \\ &= \frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta}s(\theta)\right)' \\ &= \frac{\partial}{\partial\theta'}\left(\frac{\partial}{\partial\theta}s(\theta)\right) \qquad \text{(if symmetric)} \\ &= \frac{\partial}{\partial\theta'}\left(\frac{\partial}{\partial\theta'}s(\theta)\right)' \qquad \text{(if symmetric)}. \end{aligned}$$

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

$$\frac{\partial}{\partial\theta'}h'(\theta)f(\theta) = {}_{1}h'(\theta)\,{}_{n}\left(\frac{\partial}{\partial\theta'}f(\theta)\right)_{p} + {}_{1}f'(\theta)\,{}_{n}\left(\frac{\partial}{\partial\theta'}h(\theta)\right)_{p}.$$

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

$$\frac{\partial}{\partial\rho'}f[g(\rho)] = {}_{n}\left(\frac{\partial}{\partial\theta'}f(\theta)\right)_{p}\Big|_{\theta=g(\rho)}\frac{\partial}{\partial\rho'}g(\rho)_{r}.$$

The set of nonlinear regression equations

$$y_t = f(x_t, \theta^0) + e_t \qquad t = 1, 2, \ldots, n$$

may be written in a convenient vector form

$$y = f(\theta^0) + e$$

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

$$y = \underset{n}{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_1$$

$$f(\theta) = \underset{n}{\begin{pmatrix} f(x_1, \theta) \\ f(x_2, \theta) \\ \vdots \\ f(x_n, \theta) \end{pmatrix}}_1$$

$$e = \underset{n}{\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}}_1 .$$

The sum of squared deviations

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

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

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

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

$$F(\theta) = \frac{\partial}{\partial \theta'} f(\theta);$$

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 $F(\theta^0)$ plays the same role in these linear and quadratic forms as the design matrix X in the linear regression:

$$\text{“}y\text{”} = X\beta + e.$$

The appropriate analogy is obtained by setting $\text{“}y\text{”} = y - f(\theta^0) + F(\theta^0)\theta^0$ and setting $X = F(\theta^0)$. Malinvaud (1970a, Chapter 9) terms this equation the “linear pseudo-model.” For simplicity we shall write F for the matrix $F(\theta)$ when it is evaluated at $\theta = \theta^0$:

$$F \equiv F(\theta^0).$$

Let us illustrate this notation with Example 1.

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

$$y = \begin{matrix} \\ \\ \\ \\ \\ 30 \end{matrix} \begin{pmatrix} 0.98610 \\ 1.03848 \\ 0.95482 \\ 1.04184 \\ \vdots \\ 0.50811 \\ 0.91840 \end{pmatrix}_1$$

$$f(\theta) = \begin{matrix} \\ \\ \\ \\ \\ \\ \\ 30 \end{matrix} \begin{pmatrix} \theta_1 + \theta_2 + \theta_4 e^{6.28\theta_3} \\ \theta_2 + \theta_4 e^{9.86\theta_3} \\ \theta_1 + \theta_2 + \theta_4 e^{9.11\theta_3} \\ \theta_2 + \theta_4 e^{8.43\theta_3} \\ \vdots \\ \theta_1 + \theta_2 + \theta_4 e^{0.08\theta_3} \\ \theta_2 + \theta_4 e^{6.11\theta_3} \end{pmatrix}_1 .$$

Since

$$\frac{\partial}{\partial\theta_1}f(x,\theta) = \frac{\partial}{\partial\theta_1}\left(\theta_1x_1 + \theta_2x_2 + \theta_4e^{\theta_3x_3}\right) = x_1$$

$$\frac{\partial}{\partial\theta_2}f(x,\theta) = \frac{\partial}{\partial\theta_2}\left(\theta_1x_1 + \theta_2x_2 + \theta_4e^{\theta_3x_3}\right) = x_2$$

$$\frac{\partial}{\partial\theta_3}f(x,\theta) = \frac{\partial}{\partial\theta_3}\left(\theta_1x_1 + \theta_2x_2 + \theta_4e^{\theta_3x_3}\right) = \theta_4x_3e^{\theta_3x_3}$$

$$\frac{\partial}{\partial\theta_4}f(x,\theta) = \frac{\partial}{\partial\theta_4}\left(\theta_1x_1 + \theta_2x_2 + \theta_4e^{\theta_3x_3}\right) = e^{\theta_3x_3}$$

the Jacobian of $f(\theta)$ is

$$F(\theta) = \begin{matrix}\\ \\ \\ \\ \\ \\ 30\end{matrix}\begin{pmatrix} 1 & 1 & \theta_4(6.28)e^{6.28\theta_3} & e^{6.28\theta_3} \\ 0 & 1 & \theta_4(9.86)e^{9.86\theta_3} & e^{9.86\theta_3} \\ 1 & 1 & \theta_4(9.11)e^{9.11\theta_3} & e^{9.11\theta_3} \\ 0 & 1 & \theta_4(8.43)e^{8.43\theta_3} & e^{8.43\theta_3} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \theta_4(0.08)e^{0.08\theta_3} & e^{0.08\theta_3} \\ 0 & 1 & \theta_4(6.11)e^{6.11\theta_3} & e^{6.11\theta_3} \end{pmatrix}_4 . \qquad \square$$

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 p-dimensional Euclidean space $\mathbb{R}^p$. Let θ^0 be some point in Θ.

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

$$s(\theta) = s(\theta^0) + \sum_{i=1}^{p}\left(\frac{\partial}{\partial\theta_i}s(\bar{\theta})\right)(\theta_i - \theta_i^0)$$

or, in vector notation,

$$s(\theta) = s(\theta^0) + \left(\frac{\partial}{\partial\theta}s(\bar{\theta})\right)'(\theta - \theta^0)$$

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

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

$$s(\theta) = s(\theta^0) + \sum_{i=1}^{p} \left(\frac{\partial}{\partial \theta_i} s(\theta^0) \right) (\theta_i - \theta_i^0) + \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} (\theta_i - \theta_i^0) \left(\frac{\partial^2}{\partial \theta_i \, \partial \theta_j} s(\bar{\theta}) \right) (\theta_j - \theta_j^0)$$

or, in vector notation,

$$s(\theta) = s(\theta^0) + \left(\frac{\partial}{\partial \theta} s(\theta^0) \right)' (\theta - \theta^0) + \tfrac{1}{2} (\theta - \theta^0)' \left(\frac{\partial^2}{\partial \theta \, \partial \theta'} s(\bar{\theta}) \right) (\theta - \theta^0)$$

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

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

$$f(x, \theta) = f(x, \theta^0) + \left(\frac{\partial}{\partial \theta} f(x, \theta^0) \right)' (\theta - \theta^0) + \tfrac{1}{2} (\theta - \theta^0)' \left(\frac{\partial^2}{\partial \theta \, \partial \theta'} f(x, \bar{\theta}) \right) (\theta - \theta^0)$$

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

$$f(\theta) = f(\theta^0) + \left(\frac{\partial}{\partial \theta'} f(\theta^0) \right) (\theta - \theta^0) + R(\theta - \theta^0)$$

where a typical row of R is

$$r_i' = \tfrac{1}{2} (\theta - \theta^0)' \left(\frac{\partial^2}{\partial \theta \, \partial \theta'} f(x_i, \bar{\theta}) \right) \Big|_{\bar{\theta} = \bar{\theta}(x_i, \theta)};$$

alternatively

$$f(\theta) = f(\theta^0) + F(\theta^0)(\theta - \theta^0) + R(\theta - \theta^0).$$

Using the previous formulas,

$$\begin{aligned}\frac{\partial}{\partial\theta'}\mathrm{SSE}(\theta) &= \frac{\partial}{\partial\theta'}[y-f(\theta)]'[y-f(\theta)]\\ &= [y-f(\theta)]'\frac{\partial}{\partial\theta'}[y-f(\theta)] + [y-f(\theta)]'\frac{\partial}{\partial\theta'}[y-f(\theta)]\\ &= 2[y-f(\theta)]'\left(-\frac{\partial}{\partial\theta'}f(\theta)\right)\\ &= -2[y-f(\theta)]'F(\theta).\end{aligned}$$

The least squares estimator is the value $\hat{\theta}$ that minimizes $\mathrm{SSE}(\theta)$ over the parameter space Θ. If $\mathrm{SSE}(\theta)$ is once continuously differentiable on some open set Θ^0 with $\theta \in \Theta^0 \subset \Theta$, then $\hat{\theta}$ satisfies the "normal equations"

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

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

$$y = X\beta + e$$

least squares 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. (Product rule.) Show that

$$\frac{\partial}{\partial\theta'}h'(\theta)f(\theta) = h'(\theta)\frac{\partial}{\partial\theta'}f(\theta) + f'(\theta)\frac{\partial}{\partial\theta}h(\theta)$$

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

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

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

2. (Chain rule.) Show that

$$\frac{\partial}{\partial\rho'}f[g(\rho)] = \left(\frac{\partial}{\partial\theta'}f[g(\rho)]\right)\frac{\partial}{\partial\rho'}g(\rho)$$

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^0 in the nonlinear model

$$y = f(\theta^0) + e$$

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

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

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

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

In Chapter 4 we shall show that

$$\hat{\theta} = \theta^0 + (F'F)^{-1}F'e + o_p\left(\frac{1}{\sqrt{n}}\right)$$

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

where, recall, $F = F(\theta^0) = (\partial/\partial\theta')f(\theta^0)$ is the matrix with typical row $(\partial/\partial\theta')f(x_t, \theta^0)$. 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

$$\lim_{n\to\infty} P\left[\left|\frac{X_{ijn}}{a_n}\right| > \epsilon\right] = 0$$

for any $\epsilon > 0$; $\{a_n\}$ 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 $o_p(1/n)$. Noting the similarity of the equations

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

$$s^2 = \frac{e'\left[I - F(F'F)^{-1}F'\right]e}{n-p}$$

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^0 and variance-covariance matrix $\sigma^2(F'F)^{-1}$,

$$\hat{\theta} \sim N_p\left[\theta^0, \sigma^2(F'F)^{-1}\right];$$

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

$$\frac{(n-p)s^2}{\sigma^2} \sim \chi^2(n-p);$$

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^0, s^2 converges almost surely to σ^2, $(1/n)F'(\hat{\theta})F(\hat{\theta})$ converges almost surely to a matrix Q, and $\sqrt{n}(\hat{\theta} - \theta^0)$ is asymptotically distributed as the p-variate normal with mean zero and

variance-covariance matrix $\sigma^2 Q^{-1}$,

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

The normality assumption is not needed. Let

$$\hat{Q} = \frac{1}{n} F'(\hat{\theta}) F(\hat{\theta}).$$

Following the characterization-normality approach it is natural to write

$$\hat{\theta} \doteq N_p(\theta^0, s^2 \hat{C}) \qquad \left(= N_p[\theta^0, s^2 (1/n) \hat{Q}^{-1}]\right)$$

Following the asymptotic normality approach, it is natural to write

$$\sqrt{n}(\hat{\theta} - \theta^0) \doteq N_p(0, s^2 \hat{Q}^{-1}) \qquad \left(= N_p(0, s^2 n \hat{C})\right)$$

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

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

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:

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

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^0 $(i = 1, 2, \ldots, p)$ and to test hypotheses. For example, a 95% confidence interval may be found for θ_i^0 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 $H: \theta_i^0 = \theta_i^*$ may be tested against the alternative $A: \theta_i^0 \neq \theta_i^*$ at the 5% level of significance by comparing

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

with $|t_{.025}|$ and rejecting H when $|\tilde{t}_i| > |t_{.025}|$; $\hat{c}_{ii}$ denotes the ith 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_j 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:

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

That is,

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

with

$$\begin{aligned}
a_0 &= (0,1) \\
a_1 &= (1,1).
\end{aligned}$$

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. □

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 $\mathbf{R}^{k+p}$) 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

$$s(\theta) = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \left[f(x_t, \theta) - f(x_t, \theta^0) \right]^2$$

has a unique minimum at $\theta = \theta^0$, and the second is that the matrix

$$Q = \lim_{n\to\infty} \frac{1}{n} F'(\theta^0) F(\theta^0)$$

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^0 that are often easy to spot by inspection. Although θ^0 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 $H: \theta_3^0 = 0$ or $H: \theta_4^0 = 0$ will violate the rank qualification and the identification condition, as we next show.

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

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

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

$F(\theta)$ 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 conditions that one should guard against. Let us verify that the singularity carries over to the limit. Let

$$Q_n(\theta) = \frac{1}{n}F'(\theta)F(\theta) = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{\partial}{\partial\theta}f(x_t, \theta)\right)\left(\frac{\partial}{\partial\theta}f(x_t, \theta)\right)'.$$

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

$$\lambda' Q_n(\theta)|_{\theta_3=0}\lambda = \frac{1}{n}\sum_{t=1}^{n}\left(\lambda'\frac{\partial}{\partial\theta}f(x_t, \theta)\Big|_{\theta_3=0}\right)^2 = 0.$$

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

Recall that $\{x_{3t}\}$ is independently and identically distributed according to the density $p_3(x_3)$. Since it is 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)

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

Applying these facts to the treatment group, we have

$$\lim_{n\to\infty}\frac{2}{n}\sum_{t\ \text{odd}}^{n}\left[f(x_t,\theta)-f(x_t,\theta^0)\right]^2$$
$$=\int_0^c\left[f(x,\theta)-f(x,\theta^0)\right]^2 p_3(x_3)\,dx_3\Big|_{(x_1,x_2)=(1,1)}.$$

Applying them to the control group, we have

$$\lim_{n\to\infty}\frac{2}{n}\sum_{t\ \text{even}}^{n}\left[f(x_t,\theta)-f(x_t,\theta^0)\right]^2$$
$$=\int_0^c\left[f(x,\theta)-f(x,\theta^0)\right]^2 p_3(x_3)dx_3\Big|_{(x_1,x_2)=(0,1)}.$$

Then

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\left[f(x_t,\theta^0)-f(x_t,\theta^0)\right]^2$$
$$=\frac{1}{2}\lim_{n\to\infty}\frac{2}{n}\left\{\sum_{t\ \text{odd}}^{n}\left[f(x_t,\theta)-f(x_t,\theta^0)\right]^2+\sum_{t\ \text{even}}^{n}\left[f(x_t,\theta)-f(x_t,\theta^0)\right]^2\right\}$$
$$=\frac{1}{2}\sum_{(x_1,x_2)=(0,1)}^{(1,1)}\int_0^c\left[f(x,\theta)-f(x,\theta^0)\right]^2 p_3(x_3)\,dx_3.$$

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

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

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

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

where the integral on the left is a Lebesgue-Stieltjes integral (Royden, 1968, Chapter 12; Tucker, 1967, Section 2.2). In this notation the limit can be given an integral representation

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \left[f(x_t, \theta) - f(x_t, \theta^0) \right]^2 = \int \left[f(x, \theta) - f(x, \theta^0) \right]^2 d\mu(x).$$

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

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

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

$$\begin{aligned} s(\theta) &= \int \left[f(x, \theta) - f(x, \theta^0) \right]^2 d\mu(x) \\ &= \frac{1}{2} \sum_{(x_1, x_2) = (0,1)}^{(1,1)} \int_0^c \left[f(x, \theta) - f(x, \theta^0) \right]^2 p_3(x_3)\, dx_3 \\ &= \frac{1}{2} \int_0^c \left[(\theta_2 - \theta_2^0) + \theta_4 e^{\theta_3 x} - \theta_4^0 e^{\theta_3^0 x} \right]^2 p_3(x)\, dx \\ &\quad + \frac{1}{2} \int_0^c \left[(\theta_1 - \theta_1^0) + (\theta_2 - \theta_2^0) + \theta_4 e^{\theta_3 x} - \theta_4^0 e^{\theta_3^0 x} \right]^2 p_3(x)\, dx \end{aligned}$$

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

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

Differentiating, we have

$$\theta_3 \theta_4 e^{\theta_3 x} - \theta_3^0 \theta_4^0 e^{\theta_3^0 x} = 0 \qquad 0 \leq x \leq c.$$

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

$$e^{(\theta_3 - \theta_3^0)x} = 1 \qquad 0 \le x \le c$$

which implies $\theta_3 = \theta_3^0$. We now have that

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

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

$$s(\theta) = \tfrac{1}{2}\left(\theta_2 - \theta_2^0\right)^2 + \tfrac{1}{2}\left[\left(\theta_1 - \theta_1^0\right) + \left(\theta_2 - \theta_2^0\right)\right]^2 = 0$$

which implies $\theta_1 = \theta_1^0$ and $\theta_2 = \theta_2^0$. In summary

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

As seen from Example 1, checking the identification condition and rank qualification is a tedious chore to be put to whenever 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

$$\tilde{t}_i = \frac{\theta_i - \theta_i^0}{\sqrt{s^2 \hat{c}_{ii}}}?$$

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 5000 Monte Carlo trials evaluated at the critical points of the t-distribution. The responses were generated using the inputs of Table 1