

Introduction to the Theory and Practice of Econometrics

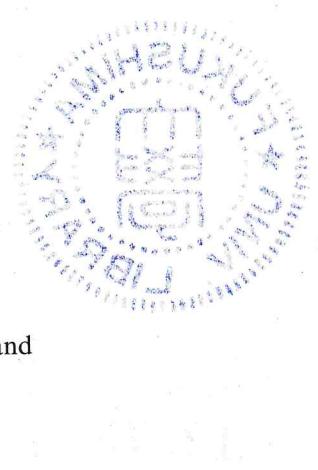
George G. Judge
University of Illinois

R. Carter Hill
University of Georgia

William Griffiths
University of New England

Helmut Lütkepohl
Universität Osnabrück

Tsoung-Chao Lee
University of Connecticut



JOHN WILEY & SONS

New York · Chichester · Brisbane · Toronto · Singapore



128753

Copyright © 1982, by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

Reproduction or translation of any part of
this work beyond that permitted by Sections
107 and 108 of the 1976 United States Copyright
Act without the permission of the copyright
owner is unlawful. Requests for permission
or further information should be addressed to
the Permissions Department, John Wiley & Sons.

Library of Congress Cataloging in Publication Data

Main entry under title:

An introduction to the theory and practice of
econometrics.

(Wiley series in probability and mathematical
statistics)

Includes bibliographical references and index.

1. Econometrics. I. Judge, George G. II. Series
HB139.158 330'.028 81-16249

AACR2

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CHAPTER 24

Nonlinear Regression Models

24.1 Introduction

Nonlinearities enter economic models in various forms. As we have seen in Chapters 6 and 7, if only the variables enter nonlinearly, the model can still be handled in the linear model framework. Furthermore, if the nonlinearity is in the parameters or in the variables and the parameters, it is sometimes possible to find a transformation to convert the considered model into a linear specification. Generally, however, this is not possible, and therefore we discuss a model of the general form

$$y_t = f(\mathbf{x}_t, \boldsymbol{\beta}) + e_t \quad (24.1.1)$$

in this chapter. In (24.1.1) \mathbf{x}_t is an $(N \times 1)$ vector of independent variables, $\boldsymbol{\beta}$ is a $(K \times 1)$ parameter vector, y_t is the dependent variable whose mean is a function of \mathbf{x}_t and $\boldsymbol{\beta}$, and e_t is a random error.

One example is the Cobb-Douglas production function

$$Q_t = \alpha L_t^{\beta_1} K_t^{\beta_2} + e_t \quad (24.1.2)$$

where Q_t is the output from a section of the economy, L_t is the labor input, and K_t is the capital input, all in period t . The parameters are α , β_1 , and β_2 . The latter two are the elasticities of Q_t with respect to L_t and K_t . Defining $y_t = Q_t$, $\mathbf{x}'_t = (L_t, K_t)$, $\boldsymbol{\beta}' = (\alpha \ \beta_1 \ \beta_2)$, and $f(\mathbf{x}_t, \boldsymbol{\beta}) = \alpha L_t^{\beta_1} K_t^{\beta_2}$, the Cobb-Douglas production function (24.1.2) has exactly the form (24.1.1). Note that, unlike linear specifications, the number of parameters, K , and the number of independent variables, N , do not necessarily coincide in nonlinear models.

Another example is a consumption function

$$C_t = \beta_1 + \beta_2 Y_t^{\beta_3} + e_t \quad (24.1.3)$$

which could arise if there is uncertainty as to the way the consumption expenditures C_t depend on the income Y_t . Why should the form

$$C_t = \beta_1 + \beta_2 Y_t^{-1} + e_t$$

be *a priori* more plausible than

$$C_t = \beta_1 + \beta_2 Y_t^{-1/2} + e_t$$

or a model with some other exponent of Y_t ? To let the data answer this question, the consumption function could be set up as in (24.1.3).

It is not unusual that parameters entering in a nonlinear way in a regression model simply reflect our uncertainty as to what model adequately represents the relationship between two or more variables. To give another example, it may not be clear whether there is a linear relationship between a dependent variable y and an independent variable x or the logarithm of x . This uncertainty can be accounted for by using a Box-Cox transformation of x , that is,

$$y_t = \beta_1 + \beta_2 \left(\frac{x_t^\lambda - 1}{\lambda} \right) + e_t \quad (24.1.4)$$

If $\lambda = 1$, we have a model

$$y_t = \alpha + \beta_2 x_t + e_t \quad (24.1.5)$$

where $\alpha = \beta_1 - \beta_2$. On the other hand, if $\lambda \rightarrow 0$ the model approaches the logarithmic specification

$$y_t = \beta_1 + \beta_2 \ln x_t + e_t \quad (24.1.6)$$

Specifying $\mathbf{x}'_t = (1 \ x_t)$, $\boldsymbol{\beta}' = (\beta_1 \ \beta_2 \ \lambda)$, and $f(\mathbf{x}_t, \boldsymbol{\beta}) = \beta_1 + \beta_2(x_t^\lambda - 1)/\lambda$ the model (24.1.4) has precisely the form (24.1.1). Models of this kind, which cannot be transformed in a linear model, are sometimes called *intrinsically nonlinear* models. If the error term is attached to the Cobb-Douglas production function in a multiplicative way, that is, if we specify

$$Q_t = \alpha L_t^{\beta_1} K_t^{\beta_2} e^{u_t} \quad (24.1.7)$$

the model is not intrinsically nonlinear because it can be linearized by taking logarithms:

$$\ln Q_t = \ln \alpha + \beta_1 \ln L_t + \beta_2 \ln K_t + u_t \quad (24.1.8)$$

We will discuss the parameter estimation for intrinsically nonlinear models like (24.1.1) in Section 24.2. Just as in the case of linear models, estimation is based on minimizing or maximizing an objective function. We consider two types of objective functions: (1) the sum of squared errors and (2) the likelihood function.

Under the standard assumptions for a linear model the minimization or maximization can be carried out by simply solving a set of linear normal equations. However, the normal equations for a nonlinear model are, in general, nonlinear in the parameters, and to solve them can be a rather difficult task. We will describe some possible solution algorithms in Section 24.3. The problem of having nonlinear normal equations is not specific to nonlinear models. As we have seen in previous chapters, the maximization of the likelihood function of a linear model may result in having to solve a nonlinear equation system. For instance, in Chapters 13 and 15 we have seen that the analysis of simultaneous equations systems and models with autocorrelated error terms can lead to such nonlinear systems. Hence it is

clear that the material in Section 24.3 is not only significant in estimating parameters of nonlinear regression models.

In Chapter 20 why and how to use prior information about the parameters that is not already contained in the model setup was discussed. This nonsample information may be given, for example, in the form of equality or inequality constraints. Of course, such information may also be available for the parameters of nonlinear regression models, and the treatment of nonsample information is discussed in Section 24.4.

If uncertainty exists regarding the nonsample information, we might want to let the data aid in deciding whether to impose the resulting constraints or not. A decision of this sort can be based on tests that are discussed in Section 24.5. Some final comments and a discussion of extensions of the considered material are given in Section 24.6.

24.2 Parameter Estimation in Nonlinear Statistical Models

To discuss the parameter estimation in a nonlinear model context let us assume that the correct model is

$$y_t = f(\mathbf{x}_t, \boldsymbol{\beta}^*) + e_t \quad (24.2.1)$$

where $\mathbf{x}_t = (x_{t1} \ x_{t2} \ \dots \ x_{tN})'$, y_t and e_t are as in (24.1.1), and $\boldsymbol{\beta}^* = (\beta_1^* \ \beta_2^* \ \dots \ \beta_K^*)'$ is the true, unknown parameter vector. The reason for using $\boldsymbol{\beta}^*$ in this chapter rather than $\boldsymbol{\beta}$, which is used in earlier chapters, is that it will be necessary later on to distinguish the true parameters from other vectors in the parameter space.

As an example of a nonlinear model let us consider

$$y_t = \beta_1^* + \beta_2^* x_{t2} + \beta_2^{*2} x_{t3} + e_t \quad (24.2.2)$$

$\mathbf{x}_t = (1 \ x_{t2} \ x_{t3})'$, $\boldsymbol{\beta}^* = (\beta_1^* \ \beta_2^*)$, and $f(\mathbf{x}_t, \boldsymbol{\beta}^*) = \beta_1^* + \beta_2^* x_{t2} + \beta_2^{*2} x_{t3}$. We may think of this specification as some sort of production function. The major reason for using it as an example arises from its simplicity, which facilitates the demonstration of the problems related to an analysis of nonlinear statistical models.

Assuming that the \mathbf{x}_t in (24.2.1) are nonstochastic and e_t fulfills the standard requirements, that is, $E[e_t] = 0$ and $E[e_t^2] = \hat{\sigma}^2$ for all t , we have $E[y_t] = f(\mathbf{x}_t, \boldsymbol{\beta}^*)$ and $\text{var}(y_t) = \hat{\sigma}^2$. Thus the mean of y_t is determined by $\boldsymbol{\beta}^*$, or, in other words, the random variable y_t depends on the unknown parameter vector. Since it will be advantageous in what follows to make this explicit, we will write $y_t(\boldsymbol{\beta}^*)$ instead of y_t . If $\boldsymbol{\beta}^*$ is replaced by any other parameter vector $\boldsymbol{\beta}$, we get a random variable $y_t(\boldsymbol{\beta})$, with a mean $f(\mathbf{x}_t, \boldsymbol{\beta})$. Written without an argument y_t denotes the observed

value of $y_t(\beta^*)$. In the following we always denote the true K -dimensional parameter vector by β^* , whereas β is just some K -dimensional vector.

Before we consider the question of how to estimate β^* , we introduce a vector notation that is often convenient. Having a sample of T observations on the dependent variable, we write

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} f(\mathbf{x}_1, \beta^*) \\ f(\mathbf{x}_2, \beta^*) \\ \vdots \\ f(\mathbf{x}_T, \beta^*) \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_T \end{bmatrix} \quad (24.2.3)$$

or, more compactly,

$$\mathbf{y} = \mathbf{f}(X, \beta^*) + \mathbf{e} \quad (24.2.4)$$

where $\mathbf{y}' = (y_1 \dots y_T)$, $\mathbf{f}(X, \beta^*) = [f(\mathbf{x}_1, \beta^*) \dots f(\mathbf{x}_T, \beta^*)]'$ and $(e_1 \dots e_T)'$ are $(T \times 1)$ vectors and $X = (\mathbf{x}_1 \dots \mathbf{x}_T)'$ is a $(T \times N)$ matrix. Let us now turn to the problem of estimating the parameter vector β^* .

24.2.1 Least Squares Estimation

For the special case of a linear model where $\mathbf{f}(X, \beta^*) = X\beta^*$ or $f(\mathbf{x}_t, \beta^*) = \mathbf{x}'_t \beta^*$, we know how to obtain good estimates. But if $\mathbf{f}(X, \beta)$ is a nonlinear function of β , the linear least squares (LS) theory is no longer valid, for the following reason.

Suppose that we choose the vector β_{LS} that minimizes the function

$$S(\beta) = [\mathbf{y} - \mathbf{f}(X, \beta)]'[\mathbf{y} - \mathbf{f}(X, \beta)] \quad (24.2.5)$$

as an estimate for β^* . Hence β_{LS} is a solution to the normal equations

$$\begin{aligned} \frac{\partial S}{\partial \beta_1} \Bigg|_{\beta_{LS}} &= -2 \sum_{t=1}^T \left[\frac{\partial f}{\partial \beta_1} \Bigg|_{\beta_{LS}} \right] [y_t - f(\mathbf{x}_t, \beta_{LS})] = 0 \\ &\vdots \\ \frac{\partial S}{\partial \beta_K} \Bigg|_{\beta_{LS}} &= -2 \sum_{t=1}^T \left[\frac{\partial f}{\partial \beta_K} \Bigg|_{\beta_{LS}} \right] [y_t - f(\mathbf{x}_t, \beta_{LS})] = 0 \end{aligned} \quad (24.2.6)$$

where $(\partial S / \partial \beta_k |_{\beta})$ is the derivative of $S(\beta)$, with respect to β_k , evaluated at β and similarly for $(\partial f / \partial \beta_k |_{\beta})$. For example, the normal equations for (24.2.2) are obtained by equating

$$\frac{\partial S}{\partial \beta_1} \Bigg|_{\beta} = -2 \sum_{t=1}^T (y_t - \beta_1 - \beta_2 x_{t2} - \beta_2^2 x_{t3}) \quad (24.2.7a)$$

$$\frac{\partial S}{\partial \beta_2} \Bigg|_{\beta} = -2 \sum_{t=1}^T (x_{t2} + 2\beta_2 x_{t3})(y_t - \beta_1 - \beta_2 x_{t2} - \beta_2^2 x_{t3}) \quad (24.2.7b)$$

to zero. In (24.2.7a) β_2 appears in a squared form, and, removing the parentheses in (24.2.7b), it is even taken to the third power. Thus the solution of the system (24.2.7) may not only be rather difficult to find, but will in general not be a linear function of the sample y . This, however, was a crucial condition used to prove the unbiasedness of the linear LS estimator (see Chapter 6). Consequently, the nonlinear LS estimator is neither linear nor unbiased, in general, and hence it is not BLUE (best linear unbiased estimator). In fact, in general, it is difficult or impossible to derive its small sample properties.

Is it reasonable under these conditions to use the least squares estimator for nonlinear models? To answer this question remember that the objective is to find a good (if not the best) estimator. Thus before we reject one possibility, we have to consider the alternatives. One of these is maximum likelihood (ML) estimation, to which we will turn now.

24.2.2 Maximum Likelihood Estimation

To use the maximum likelihood method for estimation the distribution of e has to be known except for some unknown parameters. Assuming that

$$e \sim N(\mathbf{0}, \sigma^2 I) \quad (24.2.8)$$

the likelihood function is

$$\begin{aligned} \ell(\beta, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left\{-\frac{[y - f(X, \beta)]' [y - f(X, \beta)]}{2\sigma^2}\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left\{-\frac{S(\beta)}{2\sigma^2}\right\} \end{aligned} \quad (24.2.9)$$

(compare Chapter 7). Maximizing this function is equivalent to minimizing

$$L(\beta, \sigma^2) = -2 \ln \ell(\beta, \sigma^2) = T \ln(2\pi\sigma^2) + \frac{S(\beta)}{\sigma^2} \quad (24.2.10)$$

The first-order conditions for a minimum of L are

$$\begin{aligned} \frac{\partial L}{\partial \beta_1} \Bigg|_{(\beta_{ML}, \sigma_{ML}^2)} &= \frac{1}{\sigma_{ML}^2} \left[\frac{\partial S}{\partial \beta_1} \Bigg|_{\beta_{ML}} \right] = 0 \\ &\vdots \\ \frac{\partial L}{\partial \beta_K} \Bigg|_{(\beta_{ML}, \sigma_{ML}^2)} &= \frac{1}{\sigma_{ML}^2} \left[\frac{\partial S}{\partial \beta_K} \Bigg|_{\beta_{ML}} \right] = 0 \\ \frac{\partial L}{\partial \sigma^2} \Bigg|_{(\beta_{ML}, \sigma_{ML}^2)} &= \frac{T2\pi}{2\pi\sigma_{ML}^2} - \frac{S(\beta_{ML})}{\sigma_{ML}^4} = 0 \end{aligned} \quad (24.2.11)$$

From the last equation it follows that

$$\sigma_{ML}^2 = \frac{S(\beta_{ML})}{T} \quad (24.2.12)$$

and the first K equations are only fulfilled if $(\partial S / \partial \beta_k |_{\beta}) = 0$ for $k = 1, 2, \dots, K$. However, the solution to the latter problem is not necessarily unique, since $S(\beta)$ may have several different local extrema, which is clear from (24.2.7). To see that, nevertheless, $L(\beta, \sigma^2)$ takes its minimum for that vector β_{ML} that globally minimizes $S(\beta)$, we replace σ^2 in (24.2.10) by σ_{ML}^2 given in (24.2.12) and get

$$L(\beta, \sigma_{ML}^2) = T \ln \left[2\pi \frac{S(\beta)}{T} \right] + T \quad (24.2.13)$$

This function clearly assumes its global minimum where $S(\beta)$ is minimized. Consequently, $\beta_{ML} = \beta_{LS}$, and hence the ML estimator for β^* will not be better than the LS estimator if the model disturbances are normally distributed as in (24.2.8).

Given this result, it is time to think about the adequacy of our evaluation criteria for estimator performance. It is indeed too ambitious to use small sample properties as a judgmental basis, since the wide variety of possible nonlinear specifications prevents, in general, the analytic evaluation of such properties. However, in Chapter 9 it was mentioned that ML estimators are, under general conditions, asymptotically optimal in the sense that they are consistent and have the asymptotic distribution with smallest variance. These properties do not depend on the linearity of the model. Thus we have an estimator that is good in large samples if the model errors fulfill (24.2.8). What sample can be considered large will depend on the model under investigation, however. Provided that the sample is, indeed, sufficiently large to produce estimates with properties reasonably close to those derived from the asymptotic theory, the estimator β_{ML} can be treated as being normally distributed with mean

$$E[\beta_{ML}] = \beta^* \quad (24.2.14)$$

and its variance-covariance matrix $\Sigma_{\beta_{ML}}$ can be estimated by

$$\hat{\Sigma}_{ML} = \sigma_{ML}^2 [Z(\beta_{ML})' Z(\beta_{ML})]^{-1} \quad (24.2.15)$$

where

$$Z(\beta) = \begin{bmatrix} \frac{\partial f(\mathbf{x}_1, \beta)}{\partial \beta_1} \Big|_{\beta} & \cdots & \frac{\partial f(\mathbf{x}_1, \beta)}{\partial \beta_K} \Big|_{\beta} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{x}_T, \beta)}{\partial \beta_1} \Big|_{\beta} & \cdots & \frac{\partial f(\mathbf{x}_T, \beta)}{\partial \beta_K} \Big|_{\beta} \end{bmatrix} \quad (24.2.16)$$

For instance, for (24.2.2) we have

$$Z(\beta) = \begin{bmatrix} 1 & x_{12} + 2\beta_2 x_{13} \\ 1 & x_{22} + 2\beta_2 x_{23} \\ \vdots & \vdots \\ 1 & x_{T2} + 2\beta_2 x_{T3} \end{bmatrix} \quad (24.2.17)$$

and thus

$$Z(\beta)' Z(\beta) = \begin{bmatrix} T & \sum_{t=1}^T (x_{t2} + 2\beta_2 x_{t3}) \\ \sum_{t=1}^T (x_{t2} + 2\beta_2 x_{t3}) & \sum_{t=1}^T (x_{t2} + 2\beta_2 x_{t3})^2 \end{bmatrix} \quad (24.2.18)$$

Consequently,

$$\Sigma_{ML} = \frac{\sigma_{ML}^2}{|Z(\beta_{ML})' Z(\beta_{ML})|} \begin{bmatrix} \sum_{t=1}^T (x_{t2} + 2\beta_{ML,2} x_{t3})^2 & -\sum_{t=1}^T (x_{t2} + 2\beta_{ML,2} x_{t3}) \\ -\sum_{t=1}^T (x_{t2} + 2\beta_{ML,2} x_{t3}) & T \end{bmatrix} \quad (24.2.19)$$

To demonstrate that Σ_{ML} in (24.2.15) is a reasonable estimate for the variance-covariance matrix of β_{ML} involves some mathematical manipulations, with which some readers may prefer not to bother. They may skip the next section without missing anything that is needed in later parts of the book.

24.2.3 The Asymptotic Variance-Covariance Matrix of β_{ML}

Let us define

$$\Theta = \begin{bmatrix} \beta \\ \sigma^2 \end{bmatrix} \quad (24.2.20)$$

As mentioned earlier, under fairly general conditions the ML estimator θ_{ML} is asymptotically efficient, that is, it is consistent ($\text{plim } \theta_{ML} = \theta^*$) and its asymptotic variance-covariance matrix is the Cramér-Rao lower bound given by the inverse of the information matrix

$$I(\theta^*) = -E\left[\frac{\partial^2 \ln \ell}{\partial \Theta \partial \Theta'}\Bigg|_{\theta^*}\right] = -E\left[\frac{\partial^2 \ln \ell}{\partial \theta_i \partial \theta_j}\Bigg|_{\theta^*}\right]_{i,j=1,\dots,K+1} \quad (24.2.21)$$

(Chapter 9). Let us evaluate the elements of $I(\theta^*)$.

$$\begin{aligned} \frac{\partial^2 \ln \ell}{\partial \beta \partial \beta} \Big|_{\theta^*} &= -\frac{1}{2\hat{\sigma}^2} \left[\frac{\partial^2 S}{\partial \beta_i \partial \beta_j} \Big|_{\beta^*} \right]_{i,j=1,\dots,K} \\ &= -\frac{1}{\hat{\sigma}^2} \left(\sum_{t=1}^T \left[\frac{\partial f}{\partial \beta_i} \Big|_{\beta^*} \right] \left[\frac{\partial f}{\partial \beta_j} \Big|_{\beta^*} \right] - [y_t - f(\mathbf{x}_t, \beta^*)] \left[\frac{\partial^2 f}{\partial \beta_i \partial \beta_j} \Big|_{\beta^*} \right] \right)_{i,j=1,\dots,K} \end{aligned} \quad (24.2.22)$$

Since $E[e_t] = E[y_t - f(\mathbf{x}_t, \beta^*)] = 0$ we get

$$-E\left[\frac{\partial^2 \ln \ell}{\partial \beta_i \partial \beta_j} \Big|_{\beta^*}\right] = \frac{1}{\hat{\sigma}^2} \sum_{t=1}^T \left[\frac{\partial f}{\partial \beta_i} \Big|_{\beta^*} \right] \left[\frac{\partial f}{\partial \beta_j} \Big|_{\beta^*} \right] \quad (24.2.23)$$

which is the ij th element of $(1/\hat{\sigma}^2)[Z(\beta^*)'Z(\beta^*)]$.

Furthermore,

$$\frac{\partial^2 \ln \ell}{\partial \beta_k \partial \sigma^2} \Big|_{\theta^*} = -\frac{1}{2\hat{\sigma}^4} \left[\frac{\partial S}{\partial \beta_k} \Big|_{\beta^*} \right] = \frac{1}{\hat{\sigma}^4} \sum_{t=1}^T \left[\frac{\partial f}{\partial \beta_k} \Big|_{\beta^*} \right] [y_t - f(\mathbf{x}_t, \beta^*)] \quad (24.2.24)$$

and, hence,

$$-E\left[\frac{\partial^2 \ln \ell}{\partial \beta_k \partial \sigma^2} \Big|_{\theta^*}\right] = 0 \quad (24.2.25)$$

It remains to evaluate the lower right-hand element of $I(\theta^*)$.

$$\frac{\partial^2 \ln \ell}{\partial \sigma^2 \partial \sigma^2} \Big|_{\theta^*} = \frac{T}{2\hat{\sigma}^4} - \frac{S(\beta^*)}{\hat{\sigma}^6} = \frac{T}{2\hat{\sigma}^4} - \frac{\mathbf{e}'\mathbf{e}}{\hat{\sigma}^6} \quad (24.2.26)$$

Consequently,

$$-E\left[\frac{\partial^2 \ln \ell}{\partial \sigma^2 \partial \sigma^2} \Big|_{\theta^*}\right] = \frac{-T}{2\hat{\sigma}^4} + \frac{E[\mathbf{e}'\mathbf{e}]}{\hat{\sigma}^6} = \frac{T}{2\hat{\sigma}^4} \quad (24.2.27)$$

Combining (24.2.23), (24.2.25), and (24.2.27), we can write

$$I(\theta^*) = \begin{bmatrix} \frac{1}{\hat{\sigma}^2} [Z(\beta^*)'Z(\beta^*)] & \mathbf{0} \\ \mathbf{0}' & \frac{T}{2\hat{\sigma}^4} \end{bmatrix} \quad (24.2.28)$$

Taking the inverse and replacing the true parameters by their estimates gives the estimate (24.2.15) for the variance-covariance matrix of β_{ML} .

24.2.4 The Asymptotic Distribution of β_{LS}

In many situations the distribution of the error term e_t in (24.2.1) will be unknown, in which case the ML estimation procedure is not applicable whereas least squares estimates can still be computed. What are the asymptotic properties of the LS

estimator in this case? Provided that the e_t have zero mean and are independently identically distributed with variance σ^2 , it can be shown that even if the error distribution is nonnormal the nonlinear LS estimator β_{LS} of β^* is consistent and asymptotically normally distributed under conditions that are discussed later. Assuming that the required conditions are fulfilled, let us derive the asymptotic variance-covariance matrix of β_{LS} . Since we know this matrix if the model is linear, we approximate $f(\mathbf{x}_t, \beta)$ by a linear function. The consistency of the LS estimator implies that, given a large sample, the estimated vector is likely to be close to the true parameter vector β^* . We therefore approximate $f(\mathbf{x}_t, \beta)$ in β^* by the linear term of the Taylor series expansion, which is the best linear approximation.

$$f(\mathbf{x}_t, \beta) \simeq f(\mathbf{x}_t, \beta^*) + \left[\frac{\partial f(\mathbf{x}_t, \beta)}{\partial \beta_1} \Big|_{\beta^*} \cdots \frac{\partial f(\mathbf{x}_t, \beta)}{\partial \beta_K} \Big|_{\beta^*} \right] (\beta - \beta^*) \quad (24.2.29)$$

Note that

$$\left[\frac{\partial f(\mathbf{x}_t, \beta)}{\partial \beta_1} \Big|_{\beta^*} \cdots \frac{\partial f(\mathbf{x}_t, \beta)}{\partial \beta_K} \Big|_{\beta^*} \right]$$

is the t th row of $Z(\beta^*)$ defined in (24.2.16). Thus we can write

$$\mathbf{f}(X, \beta) \simeq \mathbf{f}(X, \beta^*) + Z(\beta^*)(\beta - \beta^*) \quad (24.2.30)$$

Substituting the right-hand side for $\mathbf{f}(X, \beta)$ into $\mathbf{y} = \mathbf{f}(X, \beta) + \mathbf{e}$ gives

$$\mathbf{y} = \mathbf{f}(X, \beta^*) + Z(\beta^*)(\beta - \beta^*) + \mathbf{e}$$

or

$$\bar{\mathbf{y}}(\beta^*) = Z(\beta^*)\beta + \mathbf{e} \quad (24.2.31)$$

where

$$\bar{\mathbf{y}}(\beta) = \mathbf{y} - \mathbf{f}(X, \beta) + Z(\beta)\beta \quad (24.2.32)$$

Malinvaud (1970) calls (24.2.31) the *linear pseudomodel*. To derive this form for our example model (24.2.2), we note that in this case $Z(\beta)$, $[Z(\beta)'Z(\beta)]$, and $[Z(\beta)'Z(\beta)]^{-1}$ are given in (24.2.17), (24.2.18), and (24.2.19), respectively, and

$$\begin{aligned} \bar{\mathbf{y}}(\beta) &= \begin{bmatrix} y_1 - (\beta_1 + \beta_2 x_{12} + \beta_2^2 x_{13}) + (\beta_1 + \beta_2 x_{12} + 2\beta_2^2 x_{13}) \\ y_2 - (\beta_1 + \beta_2 x_{22} + \beta_2^2 x_{23}) + (\beta_1 + \beta_2 x_{22} + 2\beta_2^2 x_{23}) \\ \vdots \\ y_T - (\beta_1 + \beta_2 x_{T2} + \beta_2^2 x_{T3}) + (\beta_1 + \beta_2 x_{T2} + 2\beta_2^2 x_{T3}) \end{bmatrix} \\ &= \begin{bmatrix} y_1 + \beta_2^2 x_{13} \\ y_2 + \beta_2^2 x_{23} \\ \vdots \\ y_T + \beta_2^2 x_{T3} \end{bmatrix} \end{aligned} \quad (24.2.33)$$

Consequently, the linear pseudomodel corresponding to (24.2.2) is

$$\begin{bmatrix} y_1 + \beta_2^{*2} x_{13} \\ y_2 + \beta_2^{*2} x_{23} \\ \vdots \\ y_T + \beta_2^{*2} x_{T3} \end{bmatrix} = \begin{bmatrix} 1 & x_{12} + 2\beta_2^* x_{13} \\ 1 & x_{22} + 2\beta_2^* x_{23} \\ \vdots & \vdots \\ 1 & x_{T2} + 2\beta_2^* x_{T3} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_T \end{bmatrix} \quad (24.2.34)$$

Of course, since β^* is unknown, the linear pseudomodel cannot be used directly for parameter estimation. But suppose that we know $Z(\beta^*)$ and $\bar{y}(\beta^*)$. Then the linear LS estimator for β^* in (24.2.31) is

$$\hat{\beta}_L = [Z(\beta^*)' Z(\beta^*)]^{-1} Z(\beta^*) \bar{y}(\beta^*) \quad (24.2.35)$$

with variance-covariance matrix

$$\hat{\Sigma}_{\beta_L} = \hat{\sigma}^2 [Z(\beta^*)' Z(\beta^*)]^{-1} \quad (24.2.36)$$

Of course, since the model is only approximately correct, this only holds approximately. But if β_{LS} , the nonlinear LS estimator, is sufficiently close to β^* , we can use

$$\hat{\Sigma}_{\beta_{LS}} = \sigma_{LS}^2 [Z(\beta_{LS})' Z(\beta_{LS})]^{-1} \quad (24.2.37)$$

as estimate for $\Sigma_{\beta_{LS}}$. In (24.2.36) σ_{LS}^2 is an estimate for $\hat{\sigma}^2$; for instance,

$$\sigma_{LS}^2 = \frac{S(\beta_{LS})}{T - K} \quad (24.2.38)$$

The asymptotic distribution of this estimator is suggested by the LS theory for the linear pseudomodel (24.2.31) if $e \sim N(\mathbf{0}, \hat{\sigma}^2 I_T)$. Assuming that (24.2.31) is the correct model,

$$\frac{(T - K)\sigma_{LS}^2}{\hat{\sigma}^2} = \frac{e' \{I - Z(\beta^*)[Z(\beta^*)' Z(\beta^*)]^{-1} Z(\beta^*)'\} e}{\hat{\sigma}^2} \quad (24.2.39)$$

has a χ^2 distribution with $(T - K)$ degrees of freedom (compare Chapter 7). Thus this also holds asymptotically for σ_{LS}^2 computed from the nonlinear model.

We have mentioned earlier that some conditions have to be satisfied for the asymptotic theory to be valid. These are conditions for:

1. The sequence of independent variables x_t .
2. The function $f(x_t, \beta)$.
3. The errors e .

To begin with the latter, we have assumed that the e_t are independently, identically distributed with zero mean and variance $\hat{\sigma}^2$. These assumptions are sufficient if appropriately combined with conditions for the x_t and $f(x_t, \beta)$.

Concerning the properties of the latter, note that we have used the differentiability of $f(\mathbf{x}_t, \boldsymbol{\beta})$, with respect to $\boldsymbol{\beta}$, in the foregoing derivations. To establish the asymptotic normality it is convenient to assume that $f(\mathbf{x}_t, \boldsymbol{\beta})$ is continuous in both arguments and at least twice continuously differentiable with respect to $\boldsymbol{\beta}$.

Also, we have used the invertibility of $[Z(\boldsymbol{\beta}_{LS})' Z(\boldsymbol{\beta}_{LS})]$. In fact, as long as we are only concerned with the asymptotic behavior of the model it suffices to require

$$\frac{1}{T} [Z(\boldsymbol{\beta}^*)' Z(\boldsymbol{\beta}^*)] \quad (24.2.40)$$

to be nonsingular in the limit for $T \rightarrow \infty$. Although this may not be easy to verify in practice, because it depends on the sequence of independent variables \mathbf{x}_t , $t = 1, 2, \dots, T$, it is sometimes obviously violated, as we will see in Section 24.5. This problem is similar to multicollinearity in linear models. The conditions for the sequence \mathbf{x}_t , $t = 1, 2, \dots, T$ are such that it is bounded and well behaved in a certain sense as T approaches infinity. Since these properties may be difficult to establish for data that are not controlled by the investigator, they will not be spelled out here. Different sets of sufficient conditions to guarantee the consistency and asymptotic normality of nonlinear LS estimators are given in the literature (see Jennrich, 1969; Malinvaud, 1970; and Fuller, 1976).

It is not quite satisfactory to know only the asymptotic properties of our estimators, because we often do not know what sample size is sufficiently large for these asymptotic properties to hold approximately. If a large sample is not available it is difficult to feel safe about the properties of the LS or ML estimators. However, given the current state of the art, there is not much hope to derive generally valid small sample properties for nonlinear models because there are too many different possible nonlinear specifications. In the next section we will investigate how to actually compute nonlinear LS estimates.

24.3 Computing the Estimates

In the previous section we have seen that the LS estimator is a good choice for nonlinear regression models if asymptotic properties are used as the evaluation criterion and if the errors are normally distributed or if their distribution is unknown. To compute the LS estimates it is necessary to minimize the sum of squared errors. For a linear model this is relatively easy because the normal equations are linear in the unknown parameters and thus allow an analytic solution. In contrast, the following problems were identified in the previous section in minimizing the sum of squared errors of a nonlinear statistical model. First, nonlinear normal equations cannot usually be solved analytically, and second, there may be multiple solutions not all corresponding to the global

minimum of the sum of squares function. To make these problems more apparent let us consider the simple model with only one parameter

$$y_t = \beta x_{t1} + \beta^2 x_{t2} + e_t, \quad t = 1, 2, \dots, T \quad (24.3.1)$$

The normal equation of

$$S(\beta) = \sum_{t=1}^T (y_t - \beta x_{t1} - \beta^2 x_{t2})^2 \quad (24.3.2)$$

is

$$\begin{aligned} 0 &= \frac{dS}{d\beta} \Big|_\beta = -2 \sum_{t=1}^T (x_{t1} + 2\beta x_{t2})(y_t - \beta x_{t1} - \beta^2 x_{t2}) \\ &= -2 \sum_{t=1}^T y_t x_{t1} + \beta \sum_{t=1}^T (2x_{t1}^2 - 4y_t x_{t2}) + \beta^2 \sum_{t=1}^T 6x_{t1}x_{t2} + \beta^3 \sum_{t=1}^T 4x_{t2}^2 \end{aligned} \quad (24.3.3)$$

Hence we have to find the roots of a polynomial of degree 3 in order to locate the minimum of $S(\beta)$. Some possible shapes of such a polynomial are depicted in Figure 24.1, together with the corresponding sum of squared errors function.

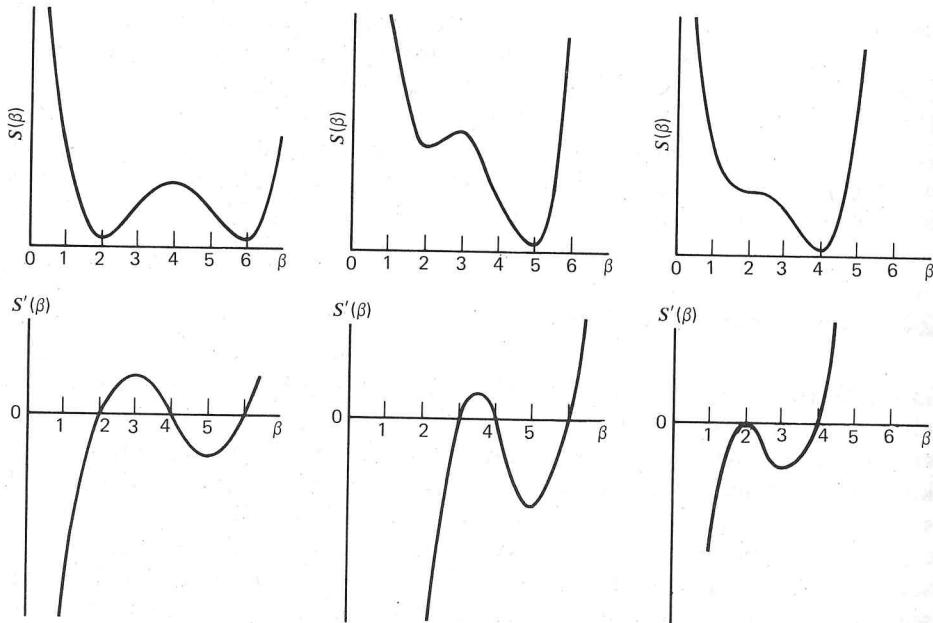


Figure 24.1 Some possible sum of squared error functions and their derivatives.

Clearly, it does not suffice to check the first-order conditions for a local extremum, but rather it is necessary to find the global minimum. This can be done by computing all solutions of the normal equations and selecting the one with a minimal corresponding sum of squared errors. Therefore, a discussion follows on how to locate solutions of the nonlinear normal equations.

24.3.1 The Gauss Method

One possible way to find the roots of the nonlinear normal equations is suggested by the discussion of the linear pseudomodel that is a linear approximation to the nonlinear model. In general, such an approximation will only be good close to the point where the derivative is evaluated. Since we do not know the true parameter vector and hence cannot approximate the nonlinear model by a linear model close to β^* , we use some other vector, for example, β_1 . Substituting this into (24.2.31), we get as linear pseudomodel

$$\bar{y}(\beta_1) = Z(\beta_1)\beta + e \quad (24.3.4)$$

The linear LS estimate of this model is

$$\beta_2 = [Z(\beta_1)'Z(\beta_1)]^{-1}Z(\beta_1)'\bar{y}(\beta_1) \quad (24.3.5)$$

Hoping that β_2 is closer than β_1 to the solution of the normal equations, we compute the next estimate replacing β_1 by β_2 in (24.3.4) and get

$$\beta_3 = [Z(\beta_2)'Z(\beta_2)]^{-1}Z(\beta_2)'\bar{y}(\beta_2) \quad (24.3.6)$$

If β_3 differs from β_2 , we can replace β_2 in (24.3.6) by β_3 to obtain β_4 , and so on. In the n th iteration we compute

$$\beta_{n+1} = [Z(\beta_n)'Z(\beta_n)]^{-1}Z(\beta_n)'\bar{y}(\beta_n) \quad (24.3.7)$$

The iteration stops if $\beta_{n+1} = \beta_n$. In this case we have indeed found a solution of the normal equations (24.2.6) because, using (24.2.32), we can write (24.3.7) as

$$\beta_{n+1} = [Z(\beta_n)'Z(\beta_n)]^{-1}Z(\beta_n)'[y - f(X, \beta_n) + Z(\beta_n)\beta_n] \quad (24.3.8)$$

or

$$[Z(\beta_n)'Z(\beta_n)]\beta_{n+1} = Z(\beta_n)'[y - f(X, \beta_n)] + [Z(\beta_n)'Z(\beta_n)]\beta_n$$

If $\beta_{n+1} = \beta_n$, this implies that

$$-2Z(\beta_n)'[y - f(X, \beta_n)] = 0 \quad (24.3.9)$$

which is precisely (24.2.6) in vector notation. This method of solving the normal equations is called the *Gauss* or *Gauss-Newton algorithm*.

To illustrate this method we use the model (24.2.2). Using (24.2.33) and (24.2.17) we get

$$Z(\beta)' \bar{y}(\beta) = \begin{bmatrix} \sum_{t=1}^T (y_t + \beta_2^2 x_{t3}) \\ \sum_{t=1}^T (y_t + \beta_2^2 x_{t3})(x_{t2} + 2\beta_2 x_{t3}) \end{bmatrix} \quad (24.3.10)$$

Left multiplying by $[Z(\beta)' Z(\beta)]^{-1}$ from (24.2.19) yields

$$\begin{aligned} & [Z(\beta)' Z(\beta)]^{-1} Z(\beta)' y(\beta) \\ &= \frac{1}{|Z(\beta)' Z(\beta)|} \begin{bmatrix} \left[\sum_{t=1}^T (x_{t2} + 2\beta_2 x_{t3})^2 \right] \left[\sum_{t=1}^T (y_t + \beta_2^2 x_{t3}) \right] \\ - \left[\sum_{t=1}^T (x_{t2} + 2\beta_2 x_{t3}) \right] \left[\sum_{t=1}^T (y_t + \beta_2^2 x_{t3})(x_{t2} + 2\beta_2 x_{t3}) \right] \\ T \left[\sum_{t=1}^T (y_t + \beta_2^2 x_{t3})(x_{t2} + 2\beta_2 x_{t3}) \right] \\ - \left[\sum_{t=1}^T (y_t + \beta_2^2 x_{t3}) \right] \left[\sum_{t=1}^T (x_{t2} + 2\beta_2 x_{t3}) \right] \end{bmatrix} \end{aligned} \quad (24.3.11)$$

Using the data in Table 24.1 we can apply this formula to compute iteratively solutions of the normal equations (24.2.7). The x_{t2} and x_{t3} in Table 24.1 are pseudo-random numbers from a uniform distribution on the unit interval, and the y_t are computed by adding a normal pseudo-random number to $\beta_1^* + \beta_2^* x_{t2} + \beta_2^{*2} x_{t3}$, where the true parameters β_1^* and β_2^* were chosen to be $\beta_1^* = \beta_2^* = 1$. That is, we have actually added the random error to $1 + x_{t2} + x_{t3}$.

As starting values for the Gauss algorithm we have arbitrarily chosen $\beta_1 = (3 \ 2)'$. The resulting iterations are shown in Table 24.2.

As we have seen, β_n being a solution of the normal equations does not necessarily mean that this vector is the global minimum of $S(\beta)$. Moreover, since the normal equations may have different solutions, it is possible that a different starting value β_1 results in a different solution. In fact, using different starting values is one way to locate the global minimum of $S(\beta)$. If different solutions of the normal equations are found, the one that corresponds to the smallest sum of squared errors has to be selected. To make sure that no local minimum is missed, a rather large number of starting values may have to be tried. This can be quite expensive in terms of computer time. Therefore, in practice the algorithm is only started at a few different

Table 24.1 Data for Example Model

<i>t</i>	<i>y_t</i>	<i>x_{t1}</i>	<i>x_{t2}</i>	<i>x_{t3}</i>
1	4.284	1.000	0.286	0.645
2	4.149	1.000	0.973	0.585
3	3.877	1.000	0.384	0.310
4	0.533	1.000	0.276	0.058
5	2.211	1.000	0.973	0.455
6	2.389	1.000	0.543	0.779
7	2.145	1.000	0.957	0.259
8	3.231	1.000	0.948	0.202
9	1.998	1.000	0.543	0.028
10	1.379	1.000	0.797	0.099
11	2.106	1.000	0.936	0.142
12	1.428	1.000	0.889	0.296
13	1.011	1.000	0.006	0.175
14	2.179	1.000	0.828	0.180
15	2.858	1.000	0.399	0.842
16	1.388	1.000	0.617	0.039
17	1.651	1.000	0.939	0.103
18	1.593	1.000	0.784	0.620
19	1.046	1.000	0.072	0.158
20	2.152	1.000	0.889	0.704

points in the parameter space, which is reasonable if the user has a rough idea of where the global minimum of $S(\beta)$ is located.

For the example model we have used two other sets of starting values, $\beta_1 = (3 -1)'$ and $\beta_1 = (1.5 0.5)'$. The resulting iterations of the Gauss algorithm are also shown in Table 24.2. For $\beta_1 = (3 -1)'$ the convergence point is a local minimum with a corresponding sum of squared errors that exceeds that obtained by using $\beta_1 = (3 2)'$ and $\beta_1 = (1.5 0.5)'$. Thus we accept

$$\beta_{LS} = \begin{bmatrix} 0.86 \\ 1.24 \end{bmatrix} \quad (24.3.12)$$

as the least squares estimate for β^* for this example. As the variance estimate we get

$$\sigma_{LS}^2 = \frac{S(\beta_{LS})}{20 - 2} = 0.89 \quad (24.3.13)$$

Furthermore,

$$[Z(\beta_{LS})' Z(\beta_{LS})]^{-1} = \begin{bmatrix} 0.26 & -0.14 \\ -0.14 & 0.10 \end{bmatrix} \quad (24.3.14)$$

Table 24.2 Iterations of the Gauss Algorithm

<i>n</i>	$\beta_{n,1}$	$\beta_{n,2}$	$S(\beta_n)$
1	3.000000	2.000000	264.3918
2	0.723481	1.404965	16.6635
3	0.837007	1.259230	16.0880
4	0.861002	1.238408	16.0818
5	0.864359	1.236040	16.0817
6	0.864740	1.235780	16.0817
7	0.864782	1.235752	16.0817
8	0.864787	1.235749	16.0817
9	0.864787	1.235749	16.0817
1	3.000000	-1.000000	25.5156
2	2.498561	-0.989894	20.4856
3	2.498566	-0.985678	20.4824
4	2.498571	-0.983903	20.4823
5	2.498574	-0.983154	20.4823
6	2.498575	-0.982837	20.4823
7	2.498576	-0.982703	20.4823
8	2.498576	-0.982646	20.4823
9	2.498576	-0.982623	20.4823
10	2.498576	-0.982612	20.4823
11	2.498576	-0.982607	20.4823
12	2.498576	-0.982605	20.4823
13	2.498576	-0.982605	20.4823
1	1.500000	0.500000	20.2951
2	1.067414	1.213585	16.6646
3	0.868351	1.233424	16.0818
4	0.865161	1.235496	16.0817
5	0.864828	1.235721	16.0817
6	0.864792	1.235746	16.0817
7	0.864788	1.235748	16.0817
8	0.864787	1.235748	16.0817
9	0.864787	1.235748	16.0817

and thus, using (24.2.37), an estimate for the variance-covariance matrix of β_{LS} is

$$\Sigma_{LS} = \sigma_{LS}^2 [Z(\beta_{LS})' Z(\beta_{LS})]^{-1} = \begin{bmatrix} 0.23 & -0.12 \\ -0.12 & 0.09 \end{bmatrix} \quad (24.3.15)$$

In practice the iteration is not necessarily carried out until β_{n+1} equals exactly β_n but is stopped once these two vectors differ by less than a prespecified small

number, such as 10^{-8} or 10^{-6} , as in our example. This number depends on the order of magnitude of the coordinates of the parameter vector. Other *termination criteria* that are often used in practice are an insignificant reduction of the sum of squared errors or reaching a maximum number of steps, that is, a maximum for the number n in (24.3.7). The latter criterion is useful because all we have demonstrated is that if the iteration converges, a solution of the normal equations is found. But there is nothing to guarantee that this will happen in a reasonable number of steps. A poor choice of the starting vector β_1 can lead the algorithm to search unsuccessfully in a region of the parameter space far away from a solution.

24.3.2 Gradient Methods

A different way to interpret the Gauss method is to regard an iteration as performing a step from some point β_n in the parameter space to a point β_{n+1} . In other words,

$$\beta_{n+1} = \beta_n + \zeta_n \quad (24.3.16)$$

where the vector ζ_n is called *step*. For the Gauss algorithm it is not difficult to determine ζ_n because (24.3.8) can be written

$$\begin{aligned} \beta_{n+1} &= \beta_n + [Z(\beta_n)'Z(\beta_n)]^{-1}Z(\beta_n)'[y - f(X, \beta_n)] \\ &= \beta_n - \frac{1}{2}[Z(\beta_n)'Z(\beta_n)]^{-1}\left[\frac{\partial S}{\partial \beta}\Big|_{\beta_n}\right] \end{aligned} \quad (24.3.17)$$

where

$$\frac{\partial S}{\partial \beta}\Big|_{\beta_n} = \left[\frac{\partial S}{\partial \beta_1}\Big|_{\beta_n} \dots \frac{\partial S}{\partial \beta_K}\Big|_{\beta_n} \right]' \quad (24.3.18)$$

is the *gradient* of $S(\beta)$ at β_n (compare Equation 24.3.9). If the iteration step is written as in (24.3.17), it is obvious that

$$\zeta_n = -\frac{1}{2}[Z(\beta_n)'Z(\beta_n)]^{-1}\left[\frac{\partial S}{\partial \beta}\Big|_{\beta_n}\right] \quad (24.3.19)$$

in (24.3.16).

In fact, (24.3.17) has the general form of the n th iteration of a so-called *gradient algorithm* for minimizing functions. The name underlines the importance of the gradient in the computation of the step ζ_n . For many gradient methods this step has the general form

$$\zeta_n = -t_n P_n \gamma_n \quad (24.3.20)$$

where γ_n is the gradient of the function to be minimized, the *objective function*; P_n is a positive definite matrix called *direction matrix*; and t_n is a positive number called the *step length*. The gradient methods differ in how they choose P_n and t_n . From (24.3.17) it follows that, for example,

$$P_n = [Z(\beta_n)' Z(\beta_n)]^{-1} \quad (24.3.21)$$

and $t_n = \frac{1}{2}$ for the Gauss method.

Other possible methods of choosing step length and step direction and their advantages and disadvantages are discussed in Judge et al. (1980). For instance, the *Newton algorithm* chooses P_n to be the inverse of the Hessian of $S(\beta)$ at β_n , that is,

$$P_n^{-1} = \mathcal{H}_n = \begin{bmatrix} \frac{\partial^2 S}{\partial \beta_1 \partial \beta_1} \Big|_{\beta_n} & \cdots & \frac{\partial^2 S}{\partial \beta_1 \partial \beta_K} \Big|_{\beta_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 S}{\partial \beta_K \partial \beta_1} \Big|_{\beta_n} & \cdots & \frac{\partial^2 S}{\partial \beta_K \partial \beta_K} \Big|_{\beta_n} \end{bmatrix} \quad (24.3.22)$$

Thus if this method is used with step length, say, $t_n = 1$, in all iterations, we choose an initial vector β_1 and compute

$$\beta_2 = \beta_1 - \mathcal{H}_1^{-1} \left[\frac{\partial S}{\partial \beta} \Big|_{\beta_1} \right]$$

and

$$\beta_3 = \beta_2 - \mathcal{H}_2^{-1} \left[\frac{\partial S}{\partial \beta} \Big|_{\beta_2} \right]$$

and so on until convergence. Again, there is no guarantee that it converges to a local minimum or even to the global minimum. For the example model (24.2.2) we have

$$\mathcal{H}_n = \begin{bmatrix} 2T & 2 \sum_{t=1}^T (x_{t2} + 2\beta_{n,2} x_{t3}) \\ 2 \sum_{t=1}^T (x_{t2} + 2\beta_{n,2} x_{t3}) & 2 \sum_{t=1}^T (2\beta_{n,1} x_{t3} + 6\beta_{n,2} x_{t2} x_{t3} + 6\beta_{n,2}^2 x_{t3}^2 + x_{t2}^2 - 2x_{t3} y_t) \end{bmatrix} \quad (24.3.23)$$

Table 24.3 Iterations of the Newton Algorithm

<i>n</i>	$\beta_{n,1}$	$\beta_{n,2}$	$S(\boldsymbol{\beta}_n)$
1	3.000000	2.000000	264.3918
2	-0.084033	1.811210	20.6328
3	0.625029	1.423940	16.5105
4	0.817259	1.272776	16.0961
5	0.862590	1.237516	16.0818
6	0.864782	1.235753	16.0817
7	0.864787	1.235748	16.0817
8	0.864787	1.235748	16.0817
1	0.000000	2.000000	29.2758
2	0.334936	1.600435	17.7382
3	0.735040	1.336953	16.1955
4	0.849677	1.247743	16.0832
5	0.864541	1.235946	16.0817
6	0.864787	1.235749	16.0817
7	0.864787	1.235748	16.0817
8	0.864787	1.235748	16.0817
1	1.500000	0.500000	20.2951
2	2.256853	0.007135	20.7735
3	2.467047	-0.436460	21.0312
4	2.316982	-0.202435	20.9467
5	2.359743	-0.320579	20.9809
6	2.354457	-0.319153	20.9805
7	2.354471	-0.319186	20.9805
8	2.354471	-0.319186	20.9805

It is left to the reader to use this result for setting up the iteration formula

$$\boldsymbol{\beta}_{n+1} = \boldsymbol{\beta}_n - \mathcal{H}_n^{-1} \left[\frac{\partial S}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}_n} \right] \quad (24.3.24)$$

for the example model. We started the Newton algorithm at different points in the parameter space and show the results in Table 24.3. In two of the three cases the global minimum is found.

The Newton algorithm, like other gradient methods, is not only applicable to minimize a sum of squared errors but can also be used to minimize other real-valued functions $H(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a parameter vector. Note that having restricted

the discussion in this section to minimizing functions is no actual constraint, because if maximization of a function $H(\theta)$ is the objective we can equivalently minimize $-H(\theta)$.

24.4 Using Nonsample Information

In Chapter 20 we considered nonsample information given as linear equality and inequality constraints for the parameters of a linear statistical model. Such outside information may also be available for the parameters of nonlinear models and is therefore relevant in a nonlinear context. Moreover, the nonsample information even for linear models may be given in the form of nonlinear equations or inequalities. Take (24.2.2) as an example. This model can be written in a linear form with a nonlinear constraint:

$$y_t = \beta_1^* + \beta_2^* x_{t2} + \beta_3^* x_{t3} + e_t \quad (24.4.1a)$$

$$\beta_3^* = \beta_2^{*2} \quad (24.4.1b)$$

Also, there are sometimes linear *and* nonlinear equality or inequality constraints for the parameters of a linear or nonlinear regression model. For instance, the stationarity constraints of a third-order autoregressive process,

$$x_t = \theta_1^* x_{t-1} + \theta_2^* x_{t-2} + \theta_3^* x_{t-3} + e_t \quad (24.4.2)$$

(see Chapter 25) can be given as inequality restrictions for the parameters as follows:

$$\begin{aligned} 1 - \theta_1^* - \theta_2^* - \theta_3^* &> 0 \\ 3 + \theta_1^* + \theta_2^* - 3\theta_3^* &> 0 \\ 1 + \theta_1^* - \theta_2^* + \theta_3^* &> 0 \\ 1 - \theta_3^{*2} + \theta_1^* \theta_3^* + \theta_2^* &> 0 \end{aligned} \quad (24.4.3)$$

(e.g., Gandolfo, 1971).

Estimating a model without known prior constraints may enable us to remain entirely in the linear model area as in (24.4.1), but we may have to pay the price of getting inefficient estimates. Moreover, the estimation results may be hard to interpret if the constraints are not imposed. In some cases we may not even be able to compute the estimates. For example, if the following logarithmic version of the CES production function

$$\ln Q_t = \beta_0 + \beta_1 \ln [\beta_2 L_t^{\beta_3} + (1 - \beta_2) K_t^{\beta_3}] + e_t \quad (24.4.4)$$

is to be analyzed, then the sum of squared errors cannot be evaluated if the term in brackets is negative. Depending on the data for L_t and K_t , we may have to keep β_2 between zero and 1. Even if we start an iterative numerical minimization algorithm with a value of β_2 in this interval, there is no guarantee that it does not produce values smaller than zero or greater than 1 for this parameter at some iteration. In such a case it may not be possible to compute the next step. These are reasons that constraints have to be considered in estimating the parameters of a regression model. To summarize, the parameter constraints can transfer us to the nonlinear world if either of the following holds:

1. The model is linear and at least one constraint is nonlinear.
2. The model is intrinsically nonlinear and all equality and inequality constraints are linear.
3. Both model and constraints are nonlinear.

We will treat equality and inequality constraints separately in the following sections. If both appear simultaneously, which is entirely possible, the methods described in the sequel can be combined. Although LS is the method of estimation throughout this section, given a model of the type (24.2.1) the discussion is not essentially different for ML estimation. An extensive list of references for the material in the following two subsections is given in Judge et al. (1980).

24.4.1 Equality Constraints

Let us assume that the model is given in the general form (24.2.1). This includes the possibility that $f(x_t, \beta)$ is a linear function of β . For simplicity we will deal with only one constraint, the multiple restriction case being analogous but notationally more involved. Furthermore, we assume that the constraint can be written in the general form

$$q(\beta) = 0 \quad (24.4.5)$$

where $q(\cdot)$ is a differentiable function. This notation means that we consider all parameters β that obey (24.4.5), since we know that the true parameter vector β^* fulfills this constraint (see Figure 24.2). For example, for (24.4.1),

$$q(\beta) = \beta_3 - \beta_2^2 \quad (24.4.6)$$

Since we know how to minimize a function without restrictions on the parameters we will try to transform the constrained minimization problem to an unconstrained one. The simplest way to do so is often to reparameterize the model such that the constraint is automatically fulfilled by all possible parameter values. For example, instead of (24.4.1), we use the model (24.2.2), that is, we

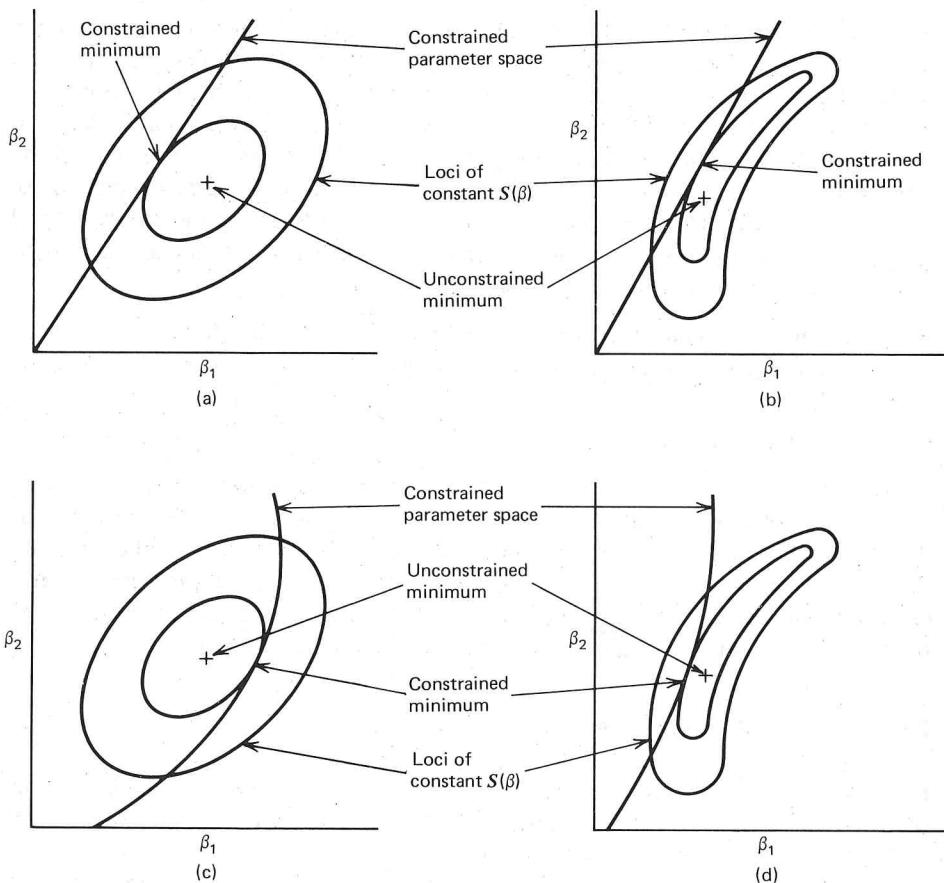


Figure 24.2 Equality constraints. (a) Linear model with linear constraints. (b) Nonlinear model with linear constraints. (c) Linear model with nonlinear constraints. (d) Nonlinear model with nonlinear constraints.

replace β_3 by β_2^2 and thereby enforce the constraint. To use the unconstrained LS estimator $\beta_{LS,2}$ for β_2^* from (24.2.2) and to estimate β_3^* by $\beta_{LS,2}^2$ is equivalent to solving the constrained LS problem implied by (24.4.1).

Unfortunately, this easy method of coping with parameter restrictions is not always applicable because it rests on the constraint being uniquely solvable for one of the parameters—in the above example, $\beta_3 = \beta_2^2$. Then β_3 is replaced by the expression on the right-hand side—that is, here, by β_2^2 . Such a transformation is not always possible. An easy example is

$$q(\beta) = \beta_1^2 + \beta_2^2 - 1 = 0 \quad (24.4.7)$$

which does not have a unique solution for either of the parameters. In such a case a *Lagrange multiplier* can be used to find the constrained minimum of $S(\beta)$. We define

$$H(\beta, \lambda) = S(\beta) + \lambda q(\beta) \quad (24.4.8)$$

where λ is the Lagrange multiplier. It can be shown that $S(\beta)$ obtains a constrained local minimum at a point where

$$\mathbf{h}(\beta, \lambda) = \left[\frac{\partial H}{\partial \beta_1} \Big|_{(\beta, \lambda)} \cdots \frac{\partial H}{\partial \beta_K} \Big|_{(\beta, \lambda)} \frac{\partial H}{\partial \lambda} \Big|_{(\beta, \lambda)} \right]' = \mathbf{0} \quad (24.4.9)$$

To use the framework of Section 24.3.2 for solving this problem we can perform an unconstrained minimization of

$$H_R(\beta, \lambda) = \mathbf{h}(\beta, \lambda)' \mathbf{h}(\beta, \lambda) \quad (24.4.10)$$

which is always nonnegative and thus has minima wherever $\mathbf{h}(\beta, \lambda) = \mathbf{0}$. Note that the last coordinate of $\mathbf{h}(\beta, \lambda)$ is

$$\frac{\partial H}{\partial \lambda} \Big|_{(\beta, \lambda)} = q(\beta) \quad (24.4.11)$$

which is necessarily zero in (24.4.9), and hence the solution fulfills the constraint.

24.4.2 Inequality Constraints

Many inequality constraints have the general form

$$q(\beta) \geq 0 \quad (24.4.12)$$

where $q(\cdot)$ is as in (24.4.5). For instance, if in (24.2.2) the nonsample information is that $\beta_1^* \geq \beta_2^*$, this can be written as

$$q(\beta^*) = \beta_1^* - \beta_2^* \geq 0 \quad (24.4.13)$$

In this case to obtain a constrained minimum of the sum of squared errors we have to minimize $S(\beta)$ over some subset of the parameter space. One possibility is to use *penalty* or *barrier functions*, as shown in Figure 24.3. These functions are added to the objective function, in our case, the sum of squared errors, to force the minimum artificially into the feasible region. A possible barrier function is

$$B(\beta) = \frac{c}{q(\beta)} \quad (24.4.14)$$

where c is a small number. Clearly, if $q(\beta) \rightarrow 0$ from above, $B(\beta) \rightarrow \infty$ (see Figure 24.3). The procedure is sometimes called the *interior point method* because it forces

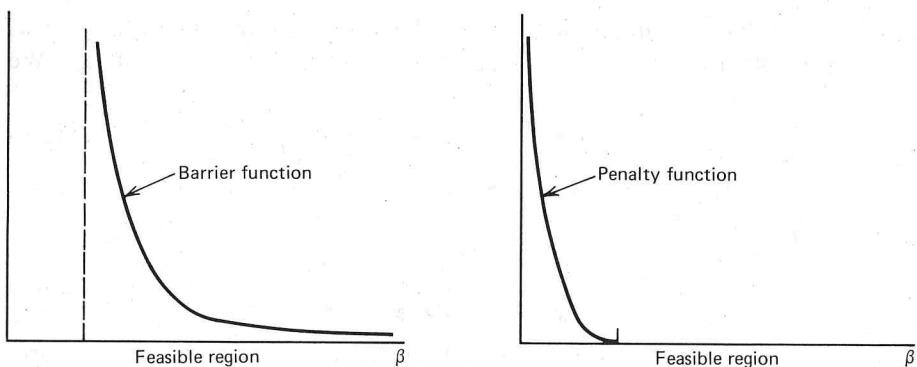


Figure 24.3 Barrier and penalty functions.

the minimization algorithm to iterate inside the feasible region, and thus points on the boundary are excluded as solutions. Since the minimum of

$$H(\beta) = S(\beta) + \frac{c}{q(\beta)} \quad (24.4.15)$$

is not necessarily identical to the constrained minimum of $S(\beta)$, it is suggested that we minimize different functions $H(\beta)$ with gradually declining values of c . If $c/q(\beta)$ is sufficiently close to zero inside the feasible region, the minimum of $H(\beta)$ and the constrained minimum of $S(\beta)$ will finally coincide, provided that the latter is not close to or on the boundary.

In this case it is advisable to use an *exterior point method*, where a *penalty function*, such as

$$\eta(\beta) = d\{\min[0, q(\beta)]\}^2 \quad (24.4.16)$$

is added to $S(\beta)$. The resulting objective function is minimized for a sequence of increasing numbers d .

Other options for inequality constrained minimization are to modify the iterative algorithm rather than the objective function or, as for equality constraints, to reparameterize the model. For instance, if $\beta_1^* > 0$ in (24.2.2), then it could be replaced by e^{x^*} , and

$$y_t = e^{x^*} + \beta_2^* x_{t2} + \beta_2^{*2} x_{t3} + e_t \quad (24.4.17)$$

can be estimated without constraints.

It is worth noting that if the true parameter vector is in the interior of the feasible region defined by some inequality constraint, then the asymptotic properties of the LS and ML estimators are not affected. However, it may still be useful to impose the restrictions, to ensure that the estimates take feasible values.

This may simplify the interpretation of the results and, as we have seen earlier, sometimes the numerical minimization requires us to account for constraints.

24.5 Confidence Regions and Hypothesis Testing

For nonlinear statistical models, as for linear models, we are sometimes interested in more than point estimates of the parameters; that is, we might like to do interval estimation and set up confidence regions. Since we have not been able to determine the small sample properties of the parameter estimators of nonlinear regression models, we clearly cannot hope to obtain precise confidence intervals. However, if the sample size is sufficiently large for the limiting distribution of the estimators discussed in Section 24.2 to be a good approximation to their actual distribution, then the asymptotic results can be used to construct confidence intervals and test hypotheses. For simplicity in the following we assume that in the nonlinear regression model (24.2.4) the error terms are normally distributed, that is, $\epsilon \sim N(\mathbf{0}, \hat{\sigma}^2 I_T)$.

We have mentioned that the nonlinear LS estimator $\hat{\beta}_{LS}$, under fairly general conditions, is consistent and has an asymptotic normal distribution with variance

$$\hat{\Sigma}_{\hat{\beta}_{LS}} = \hat{\sigma}^2 [Z(\hat{\beta}^*)' Z(\hat{\beta}^*)]^{-1} \quad (24.5.1)$$

which can be estimated by

$$\hat{\Sigma}_{LS} = \frac{S(\hat{\beta}_{LS})}{T - K} [Z(\hat{\beta}_{LS})' Z(\hat{\beta}_{LS})]^{-1} \quad (24.5.2)$$

(see (24.2.36) and (24.2.37)). Since the asymptotic variances of the components of $\hat{\beta}_{LS}$ are the diagonal elements of (24.5.1), we have approximately

$$\frac{\hat{\beta}_{LS,k} - \beta_k^*}{\sqrt{\hat{\sigma}_{LS}^2 z^{kk}}} \sim N(0, 1) \quad (24.5.3)$$

where $\hat{\beta}_{LS,k}$ is the k th coordinate of $\hat{\beta}_{LS}$, β_k^* is the k th coordinate of β^* , $\hat{\sigma}_{LS}^2$ is as defined in (24.2.38), and z^{kk} is the k th diagonal element of $[Z(\hat{\beta}_{LS})' Z(\hat{\beta}_{LS})]^{-1}$. Since $\sqrt{\hat{\sigma}_{LS}^2 z^{kk}}$ is only an estimate of the standard error of $\hat{\beta}_{LS,k}$, it might be more intuitively appealing, in analogy to the linear model case, to use the t distribution as approximation to the actual distribution of (24.5.3); that is,

$$\frac{\hat{\beta}_{LS,k} - \beta_k^*}{\sqrt{\hat{\sigma}_{LS}^2 z^{kk}}} \sim t_{T-K} \quad (24.5.4)$$

where T is the sample size as before and K is the number of parameters in the nonlinear model (compare Chapter 7). Also, however, (24.5.4) can only be justified on asymptotic grounds. But a large T implies a large number of degrees of freedom for the t distribution, which approaches the normal distribution with a growing number of degrees of freedom. Even with as little as 30 degrees of freedom there is not much difference between Student's t and the normal distribution and it is not clear which of the two is the better approximation if only a small sample is available.

24.5.1 Confidence Intervals and Confidence Regions

To establish a $(1 - \alpha)100$ percent confidence interval using (24.5.3) we proceed as in Chapter 7 and choose $z_{\alpha/2}$ from the normal distribution table such that

$$P\left\{\left|\frac{\beta_{LS,k} - \beta_k^*}{\sqrt{\sigma_{LS}^2 z^{kk}}}\right| \leq z_{\alpha/2}\right\} = 1 - \alpha \quad (24.5.5a)$$

or, equivalently,

$$P\left\{-z_{\alpha/2} \leq \frac{\beta_{LS,k} - \beta_k^*}{\sqrt{\sigma_{LS}^2 z^{kk}}} \leq z_{\alpha/2}\right\} = 1 - \alpha \quad (24.5.5b)$$

or

$$P\{\beta_{LS,k} - z_{\alpha/2}\sqrt{\sigma_{LS}^2 z^{kk}} \leq \beta_k^* \leq \beta_{LS,k} + z_{\alpha/2}\sqrt{\sigma_{LS}^2 z^{kk}}\} = 1 - \alpha \quad (24.5.5c)$$

Hence the desired $(1 - \alpha)100$ percent confidence interval for β_k^* is

$$(\beta_{LS,k} - z_{\alpha/2}\sqrt{\sigma_{LS}^2 z^{kk}}, \beta_{LS,k} + z_{\alpha/2}\sqrt{\sigma_{LS}^2 z^{kk}}) \quad (24.5.6)$$

For the example model, as 95 percent confidence intervals for β_1^* and β_2^* , using (24.3.12) to (24.3.14), we get

$$(0.86 - 1.96\sqrt{0.89 \cdot 0.26}, 0.86 + 1.96\sqrt{0.89 \cdot 0.26}) = (-0.08, 1.80)$$

$$(1.24 - 1.96\sqrt{0.89 \cdot 0.10}, 1.24 + 1.96\sqrt{0.89 \cdot 0.10}) = (0.66, 1.82)$$

If we wish to use (24.5.4) to establish a $(1 - \alpha)100$ percent confidence interval for β_k^* , the critical value $z_{\alpha/2}$ in (24.5.6) has to be replaced by the corresponding critical point from the t -distribution table.

To find simultaneous confidence regions for two or more parameters we can use the F statistic, which was given as

$$F_{(K, T-K)} = \frac{[S(\beta) - S(\beta_{LS})]/K}{S(\beta_{LS})/(T - K)} \quad (24.5.7)$$

for the linear model case in Equation 7.3.13 of Chapter 7. Remember that the reason for the quotient on the right-hand side of (24.5.7) to be distributed as an F statistic is that the numerator and the denominator divided by the population variance are independent χ^2 random variables. This, in general, does not hold for nonlinear models in small samples. However, asymptotically it is still valid as a consequence of the limiting properties of the LS estimator discussed earlier.

Even though the quotient in (24.5.7) is not exactly distributed as an F random variable, it can be used to find exact confidence contours for β^* , because

$$c = \frac{[S(\beta) - S(\beta_{LS})]/K}{S(\beta_{LS})/(T - K)} \quad (24.5.8)$$

can be transformed to

$$S(\beta) = S(\beta_{LS}) \left(1 + \frac{K}{T - K} c \right) \quad (24.5.9)$$

Consequently, we get a confidence contour for each constant c . The corresponding confidence level, however, is not known, because the actual distribution of c is unknown. An approximate confidence level can be obtained if (24.5.7) holds sufficiently approximately. In that circumstance we use that α as approximate confidence level for which $c = F_{(K, T-K, \alpha)}$. In other words, we assume that

$$P \left\{ \frac{[S(\beta) - S(\beta_{LS})]/K}{S(\beta_{LS})/(T - K)} \leq F_{(K, T-K, \alpha)} \right\} = \alpha \quad (24.5.10)$$

at least approximately. Unlike the case for linear models, the confidence contours obtained in this way will not usually be ellipsoids.

For the example model (24.2.2) we have used the data in Table 24.1 to compute the confidence contours shown in Figure 24.4.

24.5.2 Hypothesis Testing

If the sample size is sufficiently large, (24.5.7) can also be used to test hypotheses about the parameter vector β^* . For instance, to test

$$H_0: \beta^* = \alpha \quad \text{against} \quad H_a: \beta^* \neq \alpha \quad (24.5.11)$$

we compute

$$\frac{[S(\alpha) - S(\beta_{LS})]/K}{S(\beta_{LS})/(T - K)} \quad (24.5.12)$$

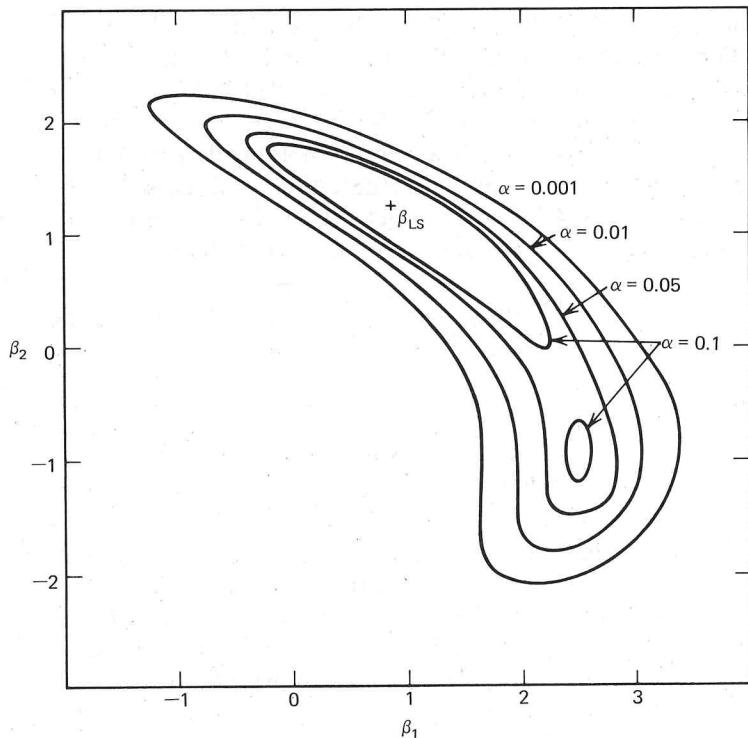


Figure 24.4 Exact confidence contours with approximate confidence levels.

and compare this number with the critical value of the F distribution that corresponds to our prespecified significance level. To test

$$H_0: \beta^* = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{against} \quad H_a: \beta^* \neq \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad (24.5.13)$$

for the example model (24.2.2) with the data in Table 24.1, we use $S([1 \ 1]') = 17.27$ to obtain a value 0.67 for (24.5.12). This value is considerably smaller than $F_{(2, 18, 0.05)} = 3.55$, and thus we cannot reject H_0 at the 0.05 level of significance. Again, this test is only asymptotically valid, and, especially in our example, the deviation of the confidence contours in Figure 24.4 from ellipses indicates that the test statistic is possibly poorly approximated by an F statistic for this small sample.

To test a hypothesis for a single parameter, the confidence intervals (24.5.6) can, of course, be used. However, here and in the foregoing it has to be kept in mind that our tests are based on the asymptotic properties of the estimators. As we have stressed in Section 24.2, the validity of the asymptotic theory is not guaranteed merely by the availability of a large sample, but also requires that some

regularity conditions hold. Although these conditions are at times difficult to verify, it is sometimes easy to see that they are *not* fulfilled. For example, consider the consumption function (24.1.3) with additive error

$$C_t = \beta_1^* + \beta_2^* Y_t^{\beta_3^*} + e_t \quad (24.5.14)$$

Here we cannot test $\beta_3^* = 0$ because this implies a model

$$C_t = \beta_1^* + \beta_2^* + e_t \quad (24.5.15)$$

with

$$Z(\beta^*)' Z(\beta^*) = \begin{bmatrix} T & T \\ T & T \end{bmatrix}$$

Hence the limit of

$$\frac{1}{T} Z(\beta^*)' Z(\beta^*) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

is singular and thereby violates a condition on which our asymptotic theory rests (see (24.2.40)).

24.6 Some Extensions

In this chapter we have discussed some aspects of analyzing nonlinear statistical models. Although the considered specifications are general enough to allow for nonlinearities in the variables and in the parameters, they are still special in a way. First, we have constrained the discussion to the single-equation case. Of course, systems of nonlinear equations do arise in practice. For examples, see Goldfeld and Quandt (1976). The problems encountered in estimation and inference are aggravated, although similar in nature to those in single-equation models. Full-information maximum likelihood estimation is one way to cope with the estimation problem theoretically. Because of the possibly high number of parameters the practical difficulties can become substantial and perhaps intractable. But if ML estimates can be found, under rather general conditions, they have the desirable properties that allow us to construct tests similar to the single-equation case. For details the reader is referred to an article by Gallant and Holly (1980).

Concerning hypothesis testing we have considered only a test for the location of the parameter vector, that is, we have discussed how to test whether the true parameters have certain values. On the other hand, we have considered much more general constraints for the parameters in Section 24.4. Since there is often uncertainty as to the adequacy of restrictions of the general type (24.4.5), it is

desirable to have tools to test such constraints. Some other tests for this purpose are discussed in Judge et al. (1980), where the topics below are also treated.

Not only may the validity of parameter constraints be open to question, but the specification of the models is also sometimes subject to uncertainty. If there is a set of possible alternative specifications, a test can be used to discriminate among them. One such test is the Cox test, which can be applied if the alternative models are not nested.

Prior nonsample knowledge about the parameters does not necessarily have the form of equality or inequality constraints as discussed in Section 24.4. If the nonsample information is given as a prior density, the Bayesian approach can be used in a nonlinear context.

We now set forth some comments on the computational methods described in Section 24.3. As we have mentioned, applying the considered algorithms can cause problems. For example, the Newton method may not converge to a minimum of the objective function. Therefore, modifications such as the *Marquardt-Levenberg algorithm* have been proposed. Another problem with both the Gauss and the Newton methods is their use of derivatives of the objective function. At times it is difficult or even impossible to obtain the required derivatives analytically, and thus methods are developed that remove this burden. Details and references are given in Judge et al. (1980).

Even though the computational difficulties can be substantial, this should not be a reason to analyze linear versions of intrinsically nonlinear models that are suggested by economic theory because in many computer centers today various packaged numerical optimization algorithms can aid in solving nonlinear estimation problems.

24.7 Problems

The following exercises are designed to lead through the steps to be carried out in an analysis of the nonlinear model

$$y_t = \beta_1^* + \beta_2^* x_{t2} + \beta_1^* \beta_2^* x_{t3} + e_t \quad (24.7.1)$$

where the e_t are normally distributed. We assume that we have a sample (y_1, y_2, \dots, y_T) of the dependent variable.

Problem 24.1

Derive the normal equations of the sum of squares function, specify the matrix $Z(\beta)$ given in (24.2.16), and write the normal equations in matrix notation.

Problem 24.2

What is the linear pseudomodel corresponding to (24.7.1)?

Problem 24.3

Write down the likelihood function, derive the information matrix, and give an analytic expression for an estimator of the variance-covariance matrix of β_{LS} .

Problem 24.4

Derive analytically the n th iteration of the Gauss algorithm.

Problem 24.5

Use the data in Table 24.1 to compute β_{LS} by applying the Gauss algorithm. Give also σ^2_{LS} , the estimate for $\hat{\sigma}^2$, and Σ_{LS} , the estimate for $\Sigma_{\beta_{LS}}$.

Problem 24.6

Establish approximate 95 percent confidence intervals for β_1^* and β_2^* .

Problem 24.7

Use the F statistic in (24.5.7) to test $\beta^* = (1.5 \ 0.5)'$ and $\beta^* = (3 \ -1)'$.

24.8 References

- Fuller, W. A. *Introduction to Statistical Time Series*. New York: John Wiley & Sons, Inc., 1976.
- Gallant, A. R., and A. Holly. Statistical Inference in an Implicit, Nonlinear, Simultaneous Equation Model in the Context of Maximum Likelihood Estimation. *Econometrica* 48: 697–720, 1980.
- Gandolfo, G. *Mathematical Methods and Models in Economic Dynamics*. Amsterdam: North-Holland, 1971.
- Goldfeld, S. M., and R. E. Quandt, editors. *Studies in Nonlinear Estimation*. Cambridge, Mass.: Ballinger, 1976.
- Jennrich, R. I. Asymptotic Properties of Nonlinear Least Squares Estimators. *Annals of Mathematical Statistics* 40: 633–643, 1969.
- Judge, G. G., W. E. Griffiths, R. C. Hill, and T. C. Lee. *The Theory and Practice of Econometrics*, New York: John Wiley & Sons, Inc., 1980.
- Malinvaud, E. *Statistical Methods of Econometrics*. Amsterdam: North-Holland, 1970.